SVEUČILIŠTE U ZAGREBU

FAKULTET STROJARSTVA I BRODOGRADNJE

DIPLOMSKI RAD

Josip Kopejtko

Zagreb, 2012.

SVEUČILIŠTE U ZAGREBU

FAKULTET STROJARSTVA I BRODOGRADNJE

DIPLOMSKI RAD

Mentor:

Doc.dr.sc. Dragutin Lisjak

Student:

Josip Kopejtko

Izjavljujem da sam ovaj rad napravio samostalno koristeći stečena znanja tijekom studija i navedenu literaturu.

Izradi ovog diplomskog rada u značajnoj su mjeri pridonijeli mnogi kolege i profesori kojima se na ovom mjestu želim zahvaliti.

Na prvom mjestu želim se zahvaliti svom mentoru Doc.dr.sc. Dragutinu Lisjaku na strpljenju, ljudskosti, a nadasve stručnosti i usmjeravanju ka korisnim raspravama bez kojih ovaj rad ne bi bio izvediv.

Hvala Prof. dr.sc. Biserki Runje na pomoći oko statističkih analiza i interpretacija.

Veliko hvala kolegi Franji Piškuliću, magistru inženjeru strojarstva, na ustupljenim podacima i pomoći oko savladavanja software-a.

Hvala kolegi Miri Hegediću, magistru inženjeru strojarstva, na savjetima,komentarima i nadahnuću u izradi ovog diplomskog rada.

Posebno hvala mojim divnim roditeljima i sestri koji su me podržavali i bili mi neprestani oslonac tijekom cijelog studija.

SADRŽAJ

POPIS SLIKA

POPIS TABLICA

SKRAĆENICE

BI – Business Inteligence

BIDS – Business Inteligence Development System

DMX – Data Mining Extension (jezik za upite na modele rudarenja)

ETL – Extract, Transform, Load (ETL alati)

HAK – Hrvatski Autoklub (korištena baza podataka)

SSAS – SQL Server Analysis Services (alati za analizu podataka)

SSMS – SQL Server Management Services (alati za upravljanje podacima)

SSIS – SQL Server Integration Services (alati za integraciju podataka)

SAŽETAK

Predviđanje potreba tržišta, rizik i neizvjesnost u procesima donošenja odluka imaju dugoročne i nesagledive posljedice za budućnost poduzeća. Kako bi te odluke bile što optimalnije, a pri tom i racionalnije, u sustav donošenja odluka uključena su i programska rješenja kao bitan proces podrške donošenju odluke.

Najvažniji segment u programskoj hijerarhiji pripada procesu otkrivanja znanja iz baze podataka. U radu je opisana struktura i značenje poslovne inteligencije kao i rudarenja podataka kao jedne od najučinkovitijih metoda za otkrivanje znanja. U skladu s time, pomoću BIDS alata opisani su algoritmi rudarenja podataka pri čemu je u praktičnom djelu rada pobliže opisan algoritam za predviđanje budućih događaja/vrijednosti temeljen na vremenskim serijama. U praktičnom djelu rada korišteni su podaci iz stvarne baze podataka te se predloženim modelima analize došlo do prihvatljivih razina predviđanja zadanih ulaznih vremenskih varijabli.

1. Osnove poslovne inteligencije

Pojava jeftinih tehnologija pohrane podataka i velika rasprostranjenost interneta, omogućila je pojedincima i organizacijama pristup velikom broju podataka. Ti su podaci često u heterogenog podrijetla i razlikuju se sadržajem i značenjem. Recimo, neki od tih podataka su komercijalne, financijske i administrativne transakcije, e-mail adrese, tekstovi i hipertekstovi, rezultati kliničkih testova, itd. Dostupnost tih podataka otvara razne mogućnosti, i postavlja pitanje: da li je moguće pretvoriti te podatke u informacije i znanja koja bi se mogla koristiti za donošenju odluka kod upravljanja poduzećima i javnom upravom? [1],[2]

Business intelligence može se definirati kao skup matematičkih modela i analiza koje koriste raspoložive podatke, te od njih stvaraju informacije i znanja koja su korisna u procesu donošenja odluka. U ovom će poglavlju biti opisani opći problemi koji se javljaju kod business intelligence, s naglaskom na povezanost s ostalim disciplinama, te će biti identificirane primarne komponente koje su tipične za business intelligence okruženje.

1.1 Učinkovite i pravovremene odluke

U složenim organizacijama odluke se donose na dnevnoj bazi. Te odluke mogu biti više ili manje važne, mogu imati dugotrajan ili kratkoročan učinak, i mogu uključivati ljude na raznim hijerarhijskim razinama. Sposobnost za donošenje odluka, bilo pojedinca ili zajednice, je jedan od primarnih faktora koji utječu na učinkovitost i konkurentnost organizacije.

Većina odluka se donosi koristeći jednostavne i intuitivne metode, koje uzimaju u obzir specifične elemente kao što su iskustvo, znanje i dostupne informacije. Takav pristup dovodi do stagnacije načina donošenja odluka, koji je neprikladan za nestabilne uvijete nametnute učestalim i brzim gospodarskim promjenama. U današnjim organizacijama, procesi odlučivanja su često presloženi i dinamični da bi se radilo na intuitivan način. Umjesto toga zahtijevaju stroži pristup temeljen na analitičkim metodama i matematičkim modelima.

Glavna uloga business intelligence sustava je pružanje alata i metoda koje omogućuju donošenje učinkovitih i pravodobnih odluka.

1.1.1 Učinkovite odluke

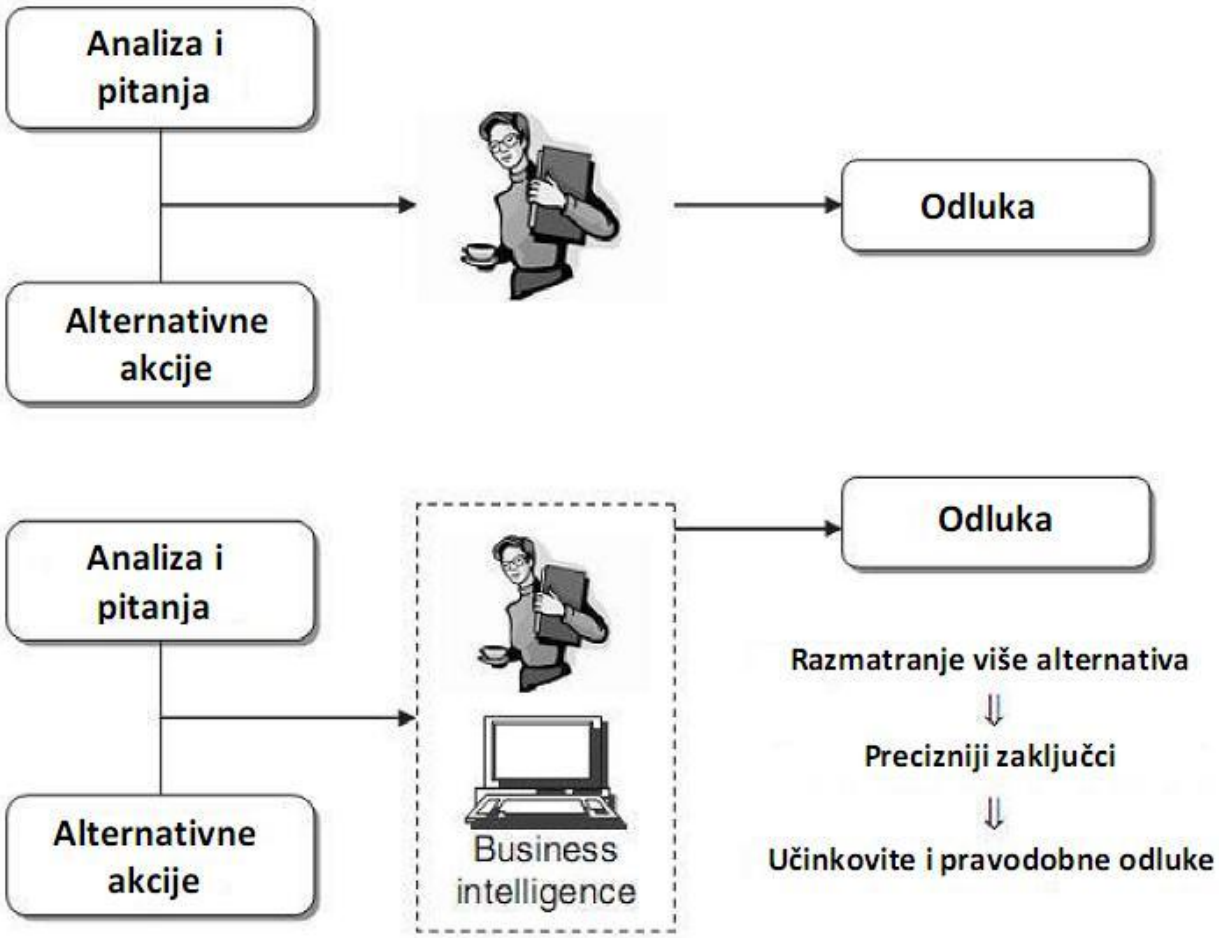
Primjena strogih analitičkih metoda omogućuje donositeljima odluka da se oslanjaju na informacije i znanja koja su pouzdanija. Rezultat toga je donošenje boljih odluka i planova

djelovanja koji omogućuju učinkovitije postizanje ciljeva. Međutim, analitičke metode zahtijevaju eksplicitan opis kriterija za procjenu alternativa i mehanizama koji reguliraju problem. Osim toga, nužan je temeljni pregled i razumijevanje temeljne logike procesa donošenja odluka.

1.1.2 Pravovremene odluke

Poduzeća posluju u gospodarskom okruženju koje karakterizira rast razine konkurencije i visoka dinamičnost. Zato je sposobnost da se brzo reagira na akcije konkurencije i na nove uvijete tržišta kritični faktor za uspjeh ili čak opstanak tvrtke.

Slika 1.1 pokazuje glavne dobiti koje organizacija može izvući uvođenjem business intelligence sustava. Kada se donosioci odluka suočavaju s problemima postavljaju si niz pitanja i razvijaju odgovarajuće analize. Nekoliko opcija se ispituje i uspoređuje, te se s obzirom na uvjete koji su uzeti u obzir, odabire najbolja. Ako se donosioci odluka mogu osloniti na business intelligence sustav, mogu se očekivati velika poboljšanja cjelokupne kvalitete procesa donošenja odluka.



Slika 1-1. Prednosti uporabe Business Intelligence sustava

Pomoću matematičkih modela i algoritama, moguće je analizirati veći broj alternativnih akcija, dolazi se do boljih zaključaka, te učinkovitijih i pravodobnih odluka. Stoga je moguće zaključiti da je glavna prednost koja proizlazi iz usvajanja business intelligence sustava povećanje učinkovitosti procesa donošenja odluka.

1.2 Podaci, informacije i znanje

U informacijskim sustavima, javnih i privatnih organizacija, nakupljaju se velika količina podataka. Ti podaci potječu dijelom od internih transakcija administrativne, logističke i komercijalne naravi, te dijelom od vanjskih izvora. Međutim, čak i ako su prikupljeni i pohranjeni na sustavan način, ne mogu se izravno koristiti za donošenje odluka. Podaci trebaju biti obrađeni odgovarajućim alatima za ekstrakciju i analitičkim metodama, sposobnim transformirati ih u informacije i znanja, koja bi se mogla koristiti za donošenje odluka.

Slijedi objašnjenje razlike između podataka, informacija i znanja.

1.2.1 Podaci

Podaci uglavnom predstavljaju kodifikaciju primarnog entiteta, kao i transakcije koje uključuju dva ili više primarnih entiteta. Primjerice, za trgovca primarni entiteti mogu biti kupci, prodajna mjesta i roba koja se prodaje, dok račun predstavlja komercijalnu transakciju.

1.2.2 Informacije

Informacije su rezultat ekstrakcije i obrade podataka, i imaju značenje onome kome su potrebne. Primjerice, voditelju prodaje maloprodajnog poduzeća, udio računa s iznosom većim od 100 novčanih jedinica u jednome tjednu, predstavlja značajnu informaciju koja može biti izvađena iz sirovih podataka.

1.2.3 Znanje

Informacija se pretvara u znanje kada se koristi za donošenje odluka i planiranje odgovarajućih akcija. Znanjem se smatra skup informacija iz nekog područja, potpomognutih iskustvom i kompetencijom donositelja odluka u rješavanju složenih problema. Analiza prodaje maloprodajnog poduzeća, može otkriti da je grupa kupaca koja živi na području, gdje je konkurent otvorio prodajno mjesto, smanjila potražnju. S vremenom će znanje prikupljeno na ovaj način dovesti do akcije koja će biti usmjerena na rješavanje problema, primjerice uvođenjem usluge besplatne dostave za kupce na tom području. Znanje može biti prikupljeno

iz podataka na pasivan način, analitičkim kriterijem predloženim od strane donositelja odluka, ili aktivnom primjenom matematičkih modela, u obliku induktivnog učenja ili optimizacije.

Postoje javna i privatna poduzeća koja su u proteklih nekoliko godina razvila formalne i sustavne mehanizme prikupljanja, pohrane i podjele znanja. Mehanizme prikupljanja sada smatraju neprocjenjivom nematerijalnom imovinom. Aktivnosti pružanja podrške u širenju znanja, kroz organizaciju, integracijom procesa donošenja odluka i usvajanjem informacijskih tehnologija obično se nazivaju upravljanje znanjem.

Očito je da business intelligence i upravljanje znanjem dijele neke sličnosti u svojim ciljevima. Glavni cilj obiju disciplina je razviti okruženje koje podupire donosioce odluka u procesu donošenja odluka i aktivnostima rješavanja složenih problema. Da bi se uočila razlika između ove dvije discipline, potrebno je primijetiti da se metode upravljanja znanjem primarno orijentiraju na obradu informacija koje su obično nestrukturirane, ponekad implicitne, i uglavnom se nalaze u dokumentima, razgovorima i iskustvima. Business intelligence se temelji na strukturiranim informacijama, najčešće kvantitativne prirode i obično organizirane u baze podataka. Međutim, ta razlika je pomalo nejasna. Primjerice, mogućnost analize e-mail adresa i internet stranica metodama rudarenja kroz tekstove, postepeno se business intelligence sustavi počinju baviti nestrukturiranim informacijama.

1.3 Uloga matematičkih modela

Business intelligence sistemi pružaju donositeljima odluka informacije i znanja izvađena iz podataka, primjenom matematičkih modela i algoritama. U nekim se slučajevima ova aktivnost može svesti na izračun postotka i grafički prikaz jednostavnih histograma, dok složenije analize zahtijevaju razvoj naprednih modela optimizacije i učenja.

Klasične znanstvene discipline, poput fizike, uvijek su posezale za matematičkim modelima da bi opisale realne sustave. Druge discipline, kao što su operacijska istraživanja, su koristile znanstvene metode i matematičke modele za proučavanje umjetnih sustava, primjerice privatnih i javnih organizacija.

Generalno gledano, usvajanje business intelligence sustava teži promicanju znanstvenog i racionalnog pristupa upravljanju poduzeća i složenih organizacija. Racionalni pristup business intelligenca može se podijeliti po glavnim koracima. Najprije se definiraju objekti analize i pokazatelji performansi koji će se koristiti za evaluaciju alternativnih mogućnosti.

Zatim se razvija matematički model koji se temelji na vezama između varijabli, parametrima i mjerama evaluacije zadanog sustava. Naposljetku se provodi what-if analiza, kojom se vrši evaluacija performansi koje ovise o varijaciji varijabla i promjenama parametara sustava.

Razvoj apstraktnog modela tjera donosioce odluka da se usredotoče na glavne značajke analiziranih sustava, što dovodi do boljeg razumijevanja danog sustava. Znanje o sustavu, koje je prikupljeno prilikom izrade matematičkog modela, moguće je jednostavno prenijeti svim pojedincima unutar organizacije. Na taj je način omogućeno bolje očuvanje znanja u odnosu na empirijski proces donošenja odluka. Matematički modeli koji su razvijeni za specifičan slučaj, vrlo su općeniti i fleksibilni da se u većini slučajeva mogu koristiti u rješavanju sličnih problema.

1.4 Business intelligence arhitektura

Arhitektura business intelligence sustava opisana je na slici 1.2, i uključuje tri glavne komponente.

1.4.1 Izvori podataka (*engl. data sources*)

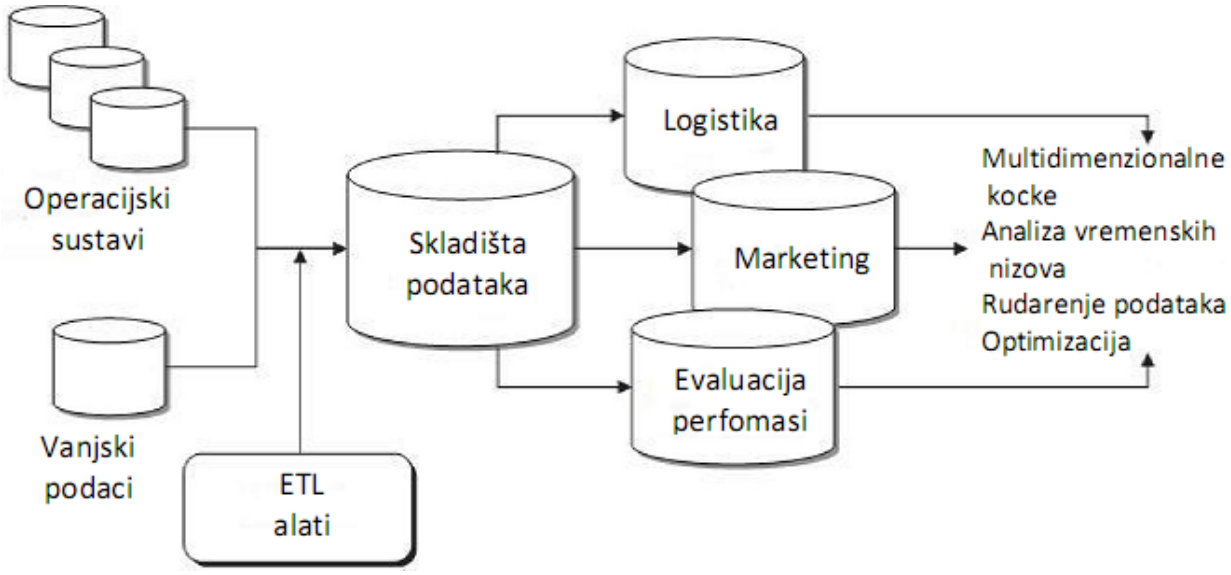
Najprije se sakupljaju i integriraju podatci, različite vrste i podrijetla, pohranjeni u raznim izvorima. Izvori se uglavnom sastoje od podataka koji pripadaju operacijskim sustavima, ali mogu sadržavati nestrukturirane dokumente kao što je elektronička pošta i podaci dobiveni od vanjskih dobavljača. Potrebni su veliki napori da se ujedine i integriraju različiti izvori podataka.

1.4.2 Skladišta podataka (engl. data warehouses and data marts)

Koriste se alati za vađenje, transformaciju i učitavanje podataka (engl. extract, transform and load, ETL), da bi se podaci iz različitih izvora sortirali u baze podataka (skladišta podataka), odnosno da bi se omogućile business intelligence analize.

1.4.3 Business intelligence metodologije (*Business intelligence methodologies*)

Nakon što su podaci pohranjeni u spremišta podataka, dostupni su matematičkim modelima i analizama namijenjenim za podršku donosiocima odluka. U business intelligence sustave, ugrađene su neke aplikacije za podršku donošenja odluka. Neke od njih su: analiza miltidimenzionalne kocke, analiza vremenskih nizova, induktivni modeli učenja za rudarenje podataka (engl. data mining), modeli za optimizaciju.



Slika 1-2. Tipična business intelligence arhitektura

Piramida na slici 1.3 prikazuje razine business intelligence sistema. Na slici 1.2 prikazane su prve dvije razine. Slijedi opis ostalih razina.

1.4.4 Istraživanje podataka (*engl. data exploration*)

Na trećoj razini piramide, nalaze se alati za provedbu pasivnih business intelligence analiza, koji se sastoje od upita (engl. query), sustava izvješćivanja (engl. reporting systems) i statističkih metoda. Nazivaju se pasivnima zato jer je donosioci odluka moraju odrediti hipoteze ili kriterije izvlačenja podataka, i onda koristiti alate za analizu. Primjerice, voditelj prodaje primijeti da su prihodi određene grupe kupaca na određenom području smanjeni. Voditelj bi možda poželio, uporabom alata za ekstrakciju i vizualizaciju, te statističkim testom provjeriti da li su njegovi zaključci popraćeni adekvatnim podacima.

1.4.5 Rudarenje podataka (*engl. data mining*)

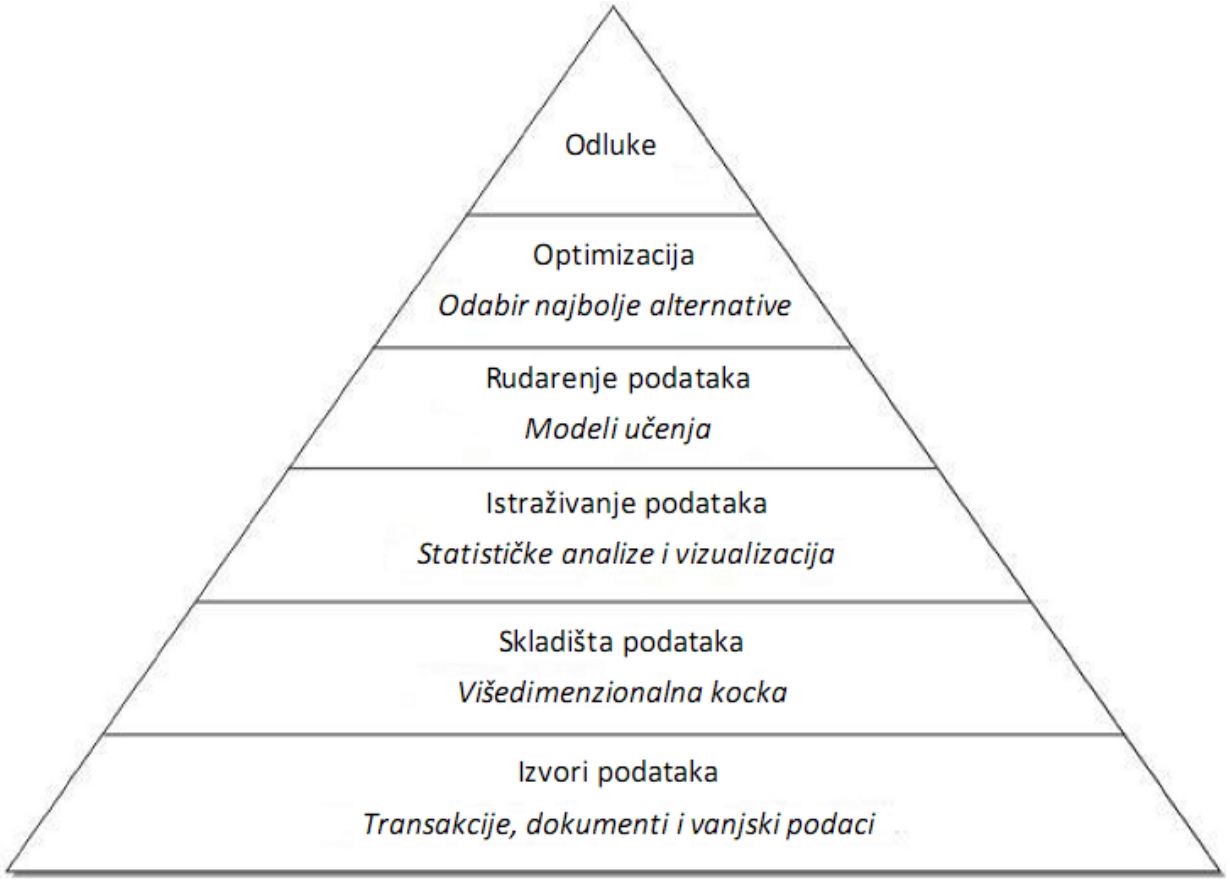
Na četvrtoj razini nalaze se aktivne business intelligence metodologije, kojima je svrha vađenje informacija i znanja iz podataka. Te metodologije uključuju matematičke modele raspoznavanja, strojno učenje i tehnike rudarenja podataka. Za razliku od alata navedenih u nižim razinama piramide, ovi su modeli aktivni, te donosioci odluka ne moraju formulirati hipotezu, koja će se kasnije provjeravati. Njihova svrha je proširiti znanje donosiocima odluka.

1.4.6 Optimizacija

Na sljedećoj razini piramide nalazi se model optimizacije koji omogućuje odabir najboljeg rješenja od mogućih alternativa. Uglavnom je skup alternativa prilično opsežan, a ponekad i beskonačan.

1.4.7 Odluka

Na vrhu piramide nalazi se odluka, koja predstavlja zaključak procesa donošenja odluka. Međutim, iako su metodologije business intelligence dostupne i uspješno usvojene, odluku moraju donijeti donosioci odluka. Donosioci odluka mogu koristiti neformalne i nestrukturirane informacije koje su im dostupne, te na temelju njih modificirati ili promijeniti odluku preporučenu od strane matematičkih modela.



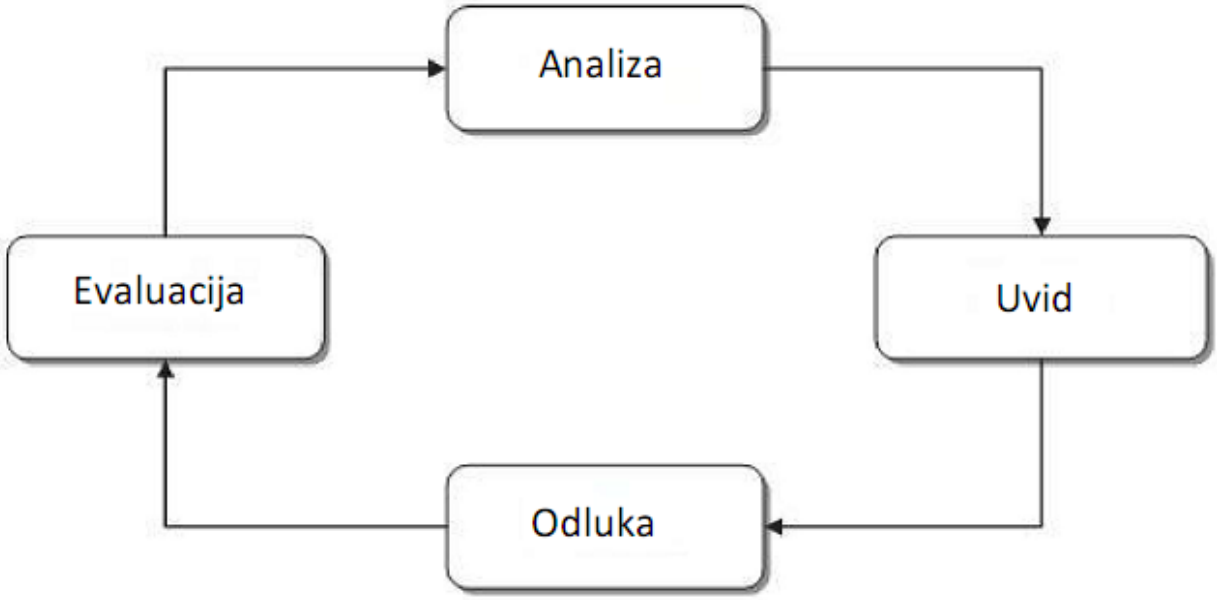
Slika 1-3. Glavne komponente business intelligence sustava

Krećući se od dna prema vrhu piramide, business intelligence sustavi nude znatno snažnije aktivne alate. Moguća je promjena stručnosti i nadležnosti zaposlenika. Na dnu piramide, potrebni su stručnjaci za informacijske sustave, koji se obično nazivaju administratorima baza podataka. Za matematičke i statističke modele, koji se nalaze u srednjim razinama piramide,

odgovorni su analitičari i stručnjaci. Za aktivnosti na vrhu piramide zaduženi su donosioci odluka.

1.5 Ciklus business intelligence analize

Svaka business intelligence analiza slijedi svoj tok ovisno o području primjene, subjektivnosti donosioca odluka, te dostupnih analitičkih metodologija. Međutim, moguće je definirati idealan ciklus business intelligence analize, kao što je prikazano na slici 1.4.



Slika 1-4. Ciklus business intelligence analize

1.5.1 Analiza

Tijekom analize prepoznaje se zadani problem i identificiraju se kritični faktori koji se doimaju najvažnijim. Primjena business intelligence metodologija omogućuje brz razvoj različitih modela istraživanja, dok se ne pronađe zadovoljavajući model.

1.5.2 Uvid

Druga faza donosiocima odluka pruža mogućnost boljeg razumijevanja danog problema. Primjerice, ako provedena analiza pokazuje da velik broj klijenata zatvara policu osiguranja nakon isteka godišnjeg ugovora, onda se u ovoj fazi identificiraju sličnosti koje ti klijenti dijele. U ovoj su fazi, informacije prikupljene u fazi analize, pretvorene u znanje. Znanje može biti stečeno intuicijom donositelja odluka, koja se temelji na njegovom iskustvu ili nestrukturiranim informacijama koje su im dostupne. S druge strane, induktivni modeli učenja pokazali su se vrlo korisnim u ovoj fazi, naročito ako su primijenjeni na strukturirane podatke.

1.5.3 Odluka

Tijekom ove faze, znanje stečeno u fazi uvida pretvara se u odluku, a zatim u akciju. Pomoću business intelligence metodologija moguće je brže izvoditi fazu analize i uvida, kako bi se postigle bolje i pravovremene odluke. Time je smanjeno vrijeme ciklusa analize, odluke, akcije i revizije, te je i donošenje odluka kvalitetnije.

1.5.4 Evaluacija

Posljednja faza business intelligence ciklusa uključuje mjerenja i evaluaciju. Smišljaju se opsežna mjerenja, koja ne uključuju samo financije, već se i ključni pokazatelji uspješnosti različitih odjela organizacije uzimaju u obzir.

1.6 Ključni faktori business intelligence projekta

Faktori koji utječu na uspjeh business intelligence projekta dijele se na tehnološke, analitičke i ljudske resurse. Neki su faktori važniji od drugih, za uspjeh projekta.

1.6.1 Tehnologija

Hardverske i softverske tehnologije bitne su faktori za razvoj business intelligence sustava poduzeća i složenih organizacija. Tijekom posljednja dva desetljeća, snaga mikroprocesora povećava se 100% svakih 18 mjeseci, a i cijena im je pada. Taj trend omogućuje primjenu naprednih algoritama koji koriste metode induktivnog učenja i modele optimizacije, vodi se računa o njihovom vremenu izvođenja. Također je moguće koristiti tehnike vizualizacije, s prikazima u realnom vremenu. Sljedeći ključni faktor je eksponencijalno povećanje kapaciteta uređaja za pohranu podataka, čija cijena također pada, što omogućuje organizacijama pohranu velikog broja podataka u sustave business intelligence. Tu je i mrežno povezivanje, odnosno ekstranet (*engl. extranet*) i intranet (*engl. intranet*), koji imaju jednu od glavnih uloga u širenju informacija i znanja unutar organizacije. Na posljetku, jednostavnost integracije hardvera i softvera različitih proizvođača, bitan je faktor koji utječe na alate za analizu podataka.

1.6.2 Analitika

Matematički modeli i analitičke tehnologije imaju ključnu ulogu u vađenju informacija i znanja iz dostupnih podataka. Sama vizualizacija podataka, pasivan je oblik podrške, ali ima važnu ulogu u olakšavanju procesa donošenja odluka. Zato se primjenjuje više naprednih

modela induktivnog učenja i optimizacije, s ciljem postizanja aktivnog oblika potpore procesu donošenja odluke.

1.6.3 Ludski potencijali

Vrijednost ljudskih potencijala čine ljudi koji djeluju unutar organizacije, bilo kao pojedinci ili timovi. Ukupna znanja koja posjeduju i dijele pojedinci čine organizacijsku kulturu. Sposobnost sakupljanja informacija i pretvaranja u akcije jedan je od glavnih faktora svake organizacije, i ima velik utjecaj na kvalitetu procesa odlučivanja. Ako je poduzeće uvelo napredni business intelligence sustav, još uvijek su potrebni radnici koji će provoditi analize, interpretirati rezultate, pronalaziti kreativna rješenja i smišljati učinkovite planove djelovanja. Organizacije, čiji su zaposlenici spremni prihvatiti promjene u načinu donošenja odluka, steći će prednost nad konkurencijom.

1.7 Razvoj business intelligence sustava

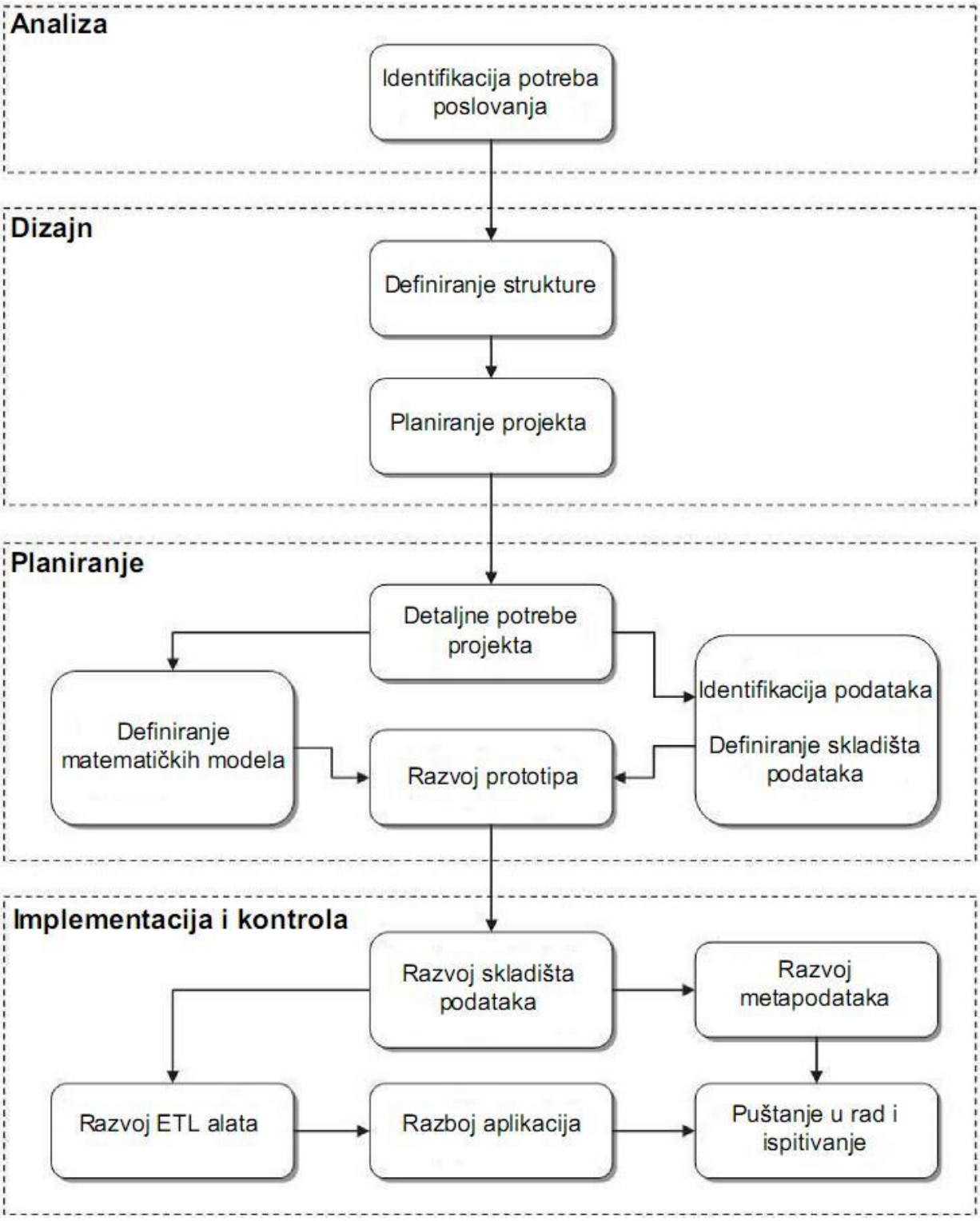
Razvoj business intelligence sustava vrlo je sličan razvoju projekta. Sličan je jer je potrebno definirati konačan cilj sustava, vrijeme i troškove razvoja sustava, te resurse koji su potrebni za izvođenje planiranih aktivnosti. Na slici 1.5 prikazan je tipičan ciklus razvoja business intelligence sustava. Razvoj business inteligence sustava može se razlikovati od prikazanog na slici. Primjerice, ako već postoje skladišta podataka sa osnovnom informacijskom strukturom, tada odgovarajuće faze sa slike 1.5 neće biti potrebne.

1.7.1 Analiza

Utvrđuju se potrebe organizacije koje su bitne za business intelligence sustav. Ta se faza obično provodi kroz niz intervjua sa zaposlenicima u organizaciji. Bitno je jasno opisati glavne ciljeve i prioritete sustava, kao i troškove i prednosti uvođenja business intelligence sustava.

1.7.2 Dizajn

Ova faza se sastoji od dvije podfaze kojima je cilj odrediti privremeni plan cjelokupne arhitekture, uzimajući u obzir razvoj u bliskoj budućnosti. Najprije se radi procjena postojećih informacijskih struktura. Da bi se prikupile sve potrebne informacije, proučava proces donošenja odluka koji će business intelligence sustav podržavati. Zatim se pomoću klasičnih metodologijama projektnog menadžmenta iznosi plan, utvrđene faze razvoja, prioriteti, očekivano vrijeme razvoja i troškovi, te potrebni resursi.



Slika 1-5. Faze razvoja business intelligence sustava

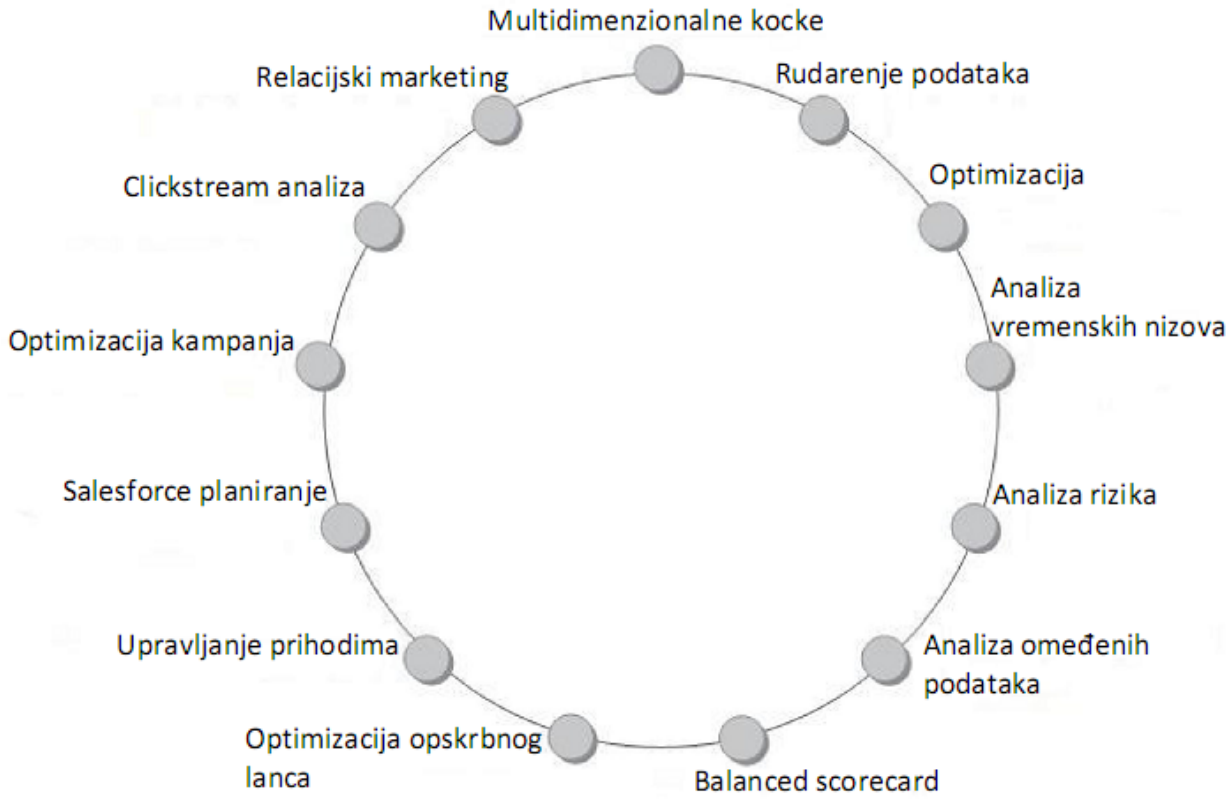
1.7.3 Planiranje

Faza planiranja uključuje podfaze, u kojima su detaljno definirane i opisane funkcije business intelligence sustava. Procjenjuju postojeći podaci i podaci iz vanjskih izvora. To omogućuje oblikovanje strukture business intelligence sustava, koja se sastoji centralnog skladišta podataka (*engl. central data warehouse*) i ako je potrebno satelitskih skladišta

podataka (*engl. satellite data mart*). Prepoznavanjem raspoloživih podataka, usvajaju se matematički modeli koje je potrebno definirati. Na taj način se osigurava dostupnost podataka za sve algoritme koji će biti korišteni za rješavanje problema. Zadnji korak ove faze je izrada prototipa, niske cijene i ograničenih mogućnosti, s ciljem otkrivanja različitosti između stvarnih potreba i onih definiranih u projekti.

1.7.4 Implementacija i kontrola

Posljednja faza se također sastoji od podfaza. Razvijaju se skladišta podataka. Ona predstavljaju informacijsku strukturu koja će business intelligence sustav snabdijevati s informacijama. Objašnjenja značenja podataka koji se nalaze u skladištima podataka, zapisuju se u dokument koji se naziva arhiva metapodataka engl. metadata). Utvrđuju se ETL procedure koje će vaditi i transformirati podatke iz primarnih izvora, te ih unositi u skladišta podataka. Razvijaju se ključne business intelligence aplikacije koje omogućuju provođenje planiranih analiza. U zadnjoj podfazi slijedi ispitivanje i korištenje sustava.



Slika 1-6. Dostupne metodologije u business intelligence sustavima

Slika 1-6 prikazuje glavne metodologije koje mogu biti korištene u business intelligence sustavima. Neke od njih mogu biti korištene u raznim područjima, dok se ostale koriste samo za određene zadatke.

2. Rudarenje podataka (*engl. data mining*)

Cilj aktivnosti uključenih u analize velikih baza podataka je vađenje korisnih informacija za podršku procesa donošenja odluka. Te se aktivnosti nazivaju: rudarenje podataka, raspoznavanje uzoraka, i strojno učenje.

Rudarenje podataka je proces istraživanja i analize skupa podataka, koji je obično velik, s ciljem pronalaženja odgovarajućih obrazaca koji se ponavljaju, kako bi se dobilo potrebno znanje. Uloga rudarenja podataka sve je važnija za teoretska istraživanja, kao i za konkretnu primjenu.

U ovom poglavlju je opisano i karakterizirano rudarenje podataka s obzirom na svrhu istraživanja i metodologije analiza. Također su objašnjena bitna svojstva ulaznih podataka, i opisan je proces rudarenja podataka. [1]

2.1 Općenito o rudarenju podataka

Aktivnosti rudarenja podataka čine iterativni procesi usmjereni na analizu velikih baza podataka, kako bi se izvadile informacije i znanja koja mogu biti korisna u donošenju odluka. Proces analize je iterativan i ima različite faze koje mogu slati povratne informacije i izvoditi naknadne revizije. Za oblikovanje takvih procesa obično je potrebna suradnja stručnjaka na području u kojem se primjenjuje i analitičara podataka koji koriste matematičke modele za induktivno učenje. Praksa pokazuje da izvođenje rudarenja podataka zahtijeva česte intervencije analitičara tijekom različitih faza, te stoga se ne može lako automatizirati. Nužno je da je izvađeno znanje točno, odnosno da je utemeljeno na podacima, i ne smije dovoditi do pogrešnih zaključaka.

Pojam rudarenje podataka odnosi se na sve procese prikupljanja i analiziranja podataka, razvoj modela induktivnog učenja, usvajanje odluka i akcije koje su rezultat stečenog znanja. Pojam teorija matematičkog učenja (*engl. mathematical learning theory*) odnosi se na niz matematičkih modela i metoda koje se mogu naći u jezgri svake analize rudarenja podataka koja se koristi za stvaranje novih znanja.

Proces rudarenja podataka temelji se na metodama induktivnog učenja, čija je glavna svrha izvođenje općih pravila počevši od dostupnih primjera koji se sastoje od zabilježenih promatranja iz prošlosti sačuvanih u jednoj ili više baza podataka. Drugim riječima, svrha analiza rudarenja podataka je izvući neke zaključke iz opažanja iz prošlosti i generalizirati ih

na što je moguće precizniji način. Modeli i obrasci koji se utvrde na taj način mogu biti različitog oblika, primjerice, linearne jednadžbe, skup if-then-else pravila, klasteri i grafovi.

Nadalje, rudarenje podataka ovisi o postupcima prikupljanja zapažanja i unosom istih u bazu podataka, iako primarna svrha pohrane tih podataka nisu analize rudarenja podataka. Primjerice, telefonske kompanije zapise o pozivima pohranjuju za administrativnu upotrebu, dok se kasnije oni mogu koristiti za izvođenje analiza rudarenja podataka. Procedura prikupljanja podataka ne ovisi, niti je u skladu s ciljevima rudarenja podataka, tako da se to prikupljanje razlikuje od prikupljanja podataka koje se provodi za uzimanje uzoraka, po unaprijed određenoj shemi.

Aktivnosti rudarenja podataka mogu se podijeliti u dvije skupine, u klasu s glavnim ciljem analize: interpretacija i predviđanje.

2.1.1 Interpretacija

Cilj interpretacije je identificirati obrasce podataka koji se ponavljaju i zatim ih izraziti pravilima i kriterijima koje će stručnjaci na području primjene lako razumjeti. Pravila moraju biti originalna i netrivijalna kako bi se povećala razina znanja i razumijevanja promatranog sustava. Primjerice, za maloprodajnu industriju, moglo bi biti korisno klasificirati kupce, koji posjeduju potrošačke kartice (engl. loyalty card), ovisno o profilu kupnje. Na taj se način stvara segment koji bi mogao biti koristan za pronalaženju marketinške niše i usmjeravanje budućih marketinških kampanja.

2.1.2 Predviđanje

Svrha predviđanja je predvidjeti koju će vrijednost slučajna varijabla poprimiti u budućnosti ili procijeniti vjerojatnost budućih događaja. Primjerice, telefonske kompanije mogu razviti analizu rudarenja podataka za procjenu vjerojatnosti prelaska klijenata kod konkurencije. Većina tehnika rudarenja podataka vrši predviđanja na temelju vrijednosti varijabli entiteta u bazi podataka. Primjerice, model rudarenja podatak može ukazivati na to da vjerojatnost prelaska klijenata kod konkurencije ovisi o značajkama kao što su dob, trajanje ugovora i udio poziva na mreže konkurencije.

Ponekad, model koji je razvijen u svrhu predviđanja može također biti koristan za interpretaciju.

2.2 Modeli i metode rudarenja podataka

Postoji nekoliko metoda učenja za izvođenje različitih zadataka rudarenja podataka. Velik broj tehnika potječe s područja računalne znanosti, kao što je klasifikacija stabala ili asocijativna pravila, i nazivaju se strojno učenje (*engl. machine learning*) ili otkrivanje znanja (*engl. knowledge discovery*). Unutar ove vrste tehnika u većini slučajeva prevladava empirijski pristup. Ostale metode spadaju u viševarijantnu statistiku, kao što je regresija ili Bayesianovi klasifikatori, i često su parametarske i teoretski više utemeljene. U zadnje vrijeme koriste se matematičke metode učenja, kao što je statistička teorija učenja (*engl. statistical learning theory*), koje se temelje na čvrstim teoretskim osnovama i uključuju više disciplina, među kojima se nalaze teorija vjerojatnosti, optimizacije i statistike.

Bez obzira na specifične metode učenja koje se žele usvojiti, postoje koraci za razvoj modela rudarenja podataka:

- odabir klase modela koji će se koristiti za učenje iz prošlosti i specifičnog oblika predstavljanja obrazaca
- definiranje mjera za evaluaciju učinkovitosti i točnosti odabranog modela
- oblikovanje računalnog algoritma, za generiranje modela, optimizirajući mjere evaluacije

2.3 Rudarenje podataka, klasična statistika i OLAP

Projekti rudarenja podataka razlikuju se po klasičnoj statistici i OLAP analizama. Razlike su prikazana u tablici 2-1, s obzirom na primjer.

Tablica 1. Razlike između OLAP, statistike i rudarenja podataka

OLAP	**Statistika**	**Rudarenje podataka**
ekstrakcija detalja i agregiranje podataka	provjera hipoteze koju je kreirao analitičar	identifikacija obrazaca podataka
informacija	provjera	znanje
raspodjela prihoda potraživača zajma za kuću	analiza varijance prihoda potraživača	karakterizacija kuće i predviđanje budućih potraživača

Glavna razlika je u aktivnoj orijentaciji koju nude modeli induktivnog učenja, za razliku od statističkih tehnika i OLAP koje su pasivnije. Statističke analize donositelji odlika formuliraju hipotetski, te one trebaju biti potvrđene na temelju uzoraka. Slično tome, na temelju intuicije donositelja odluka OLAP analize temelje ekstrakciju, izvještavanje i kriterije

vizualizacije. Statističke metoda i alati za navigaciju kroz kocke podataka koriste elemente za potvrdu ili odbacivanje formuliranih hipoteza, tijek analize je top-down. Suprotno tome, modeli učenja svojom aktivnom ulogom generiraju predviđanja i interpretacije koje predstavljaju nova znanja koja su dostupna korisnicima. U ovom slučaju je tijek analize bottom-up. Kada je u pitanju velika količina podataka, tada modeli koji su sposobni aktivno djelovati imaju bitnu ulogu, jer je korisnicima teško formulirati smislene i dobro utemeljene hipoteze.

2.4 Primjene rudarenja podataka

Metode rudarenja podataka mogu se primijeniti na razna područja, od marketinga i proizvodnih procesa do istraživanja rizičnih faktora medicinskih dijagnoza, zatim od procijene učinkovitosti novih lijekova do otkrivanja prijevara.

2.4.1 Relacijski marketing

Aplikacije rudarenja podataka na području marketinga znatno su pridonijele popularnosti ovih metodologija. Neke od bitnih aplikacija ovog područja su:

- identifikacija segmenata kupaca koji će reagirati na ciljane marketinške kampanje
- identifikacija segmenata kupaca za zadržavanje kampanje
- predviđanje stope pozitivne reakcije na marketinške kampanje
- interpretacija i razumijevanje ponašanje kupaca u kupovini
- analiza proizvoda koje kupac zajedno kupuje prilikom jedne kupovine, poznata kao analiza potrošačke košarice (engl. market basket analysis)

2.4.2 Otkrivanje prijevare

Prijevare se mogu javiti u industrijama kao što su telefonija, osiguranja (lažne potražnje) i bankarstvo (ilegalna uporaba kreditnih kartica i čekova; ilegalne novčane transakcije).

2.4.3 Procjena rizika

Svrha analize rizika je procjena rizika povezanog s budućim odlukama. Primjerice, koristeći dostupna zapažanja iz prošlosti, banka može razviti model za predviđanje da bi utvrdila da li je prikladno dati novčani zajam, ovisno o karakteristikama podnositelja zahtjeva.

2.4.4 Rudarenje teksta

Rudarenje podataka se može primijeniti na različite vrste tekstova, koji predstavljaju nestrukturirane podatke, kako bi se klasificirali članci, knjige, dokumenti, elektronička pošta i

web stranice. Najbolji primjeri su web tražilice. Ostale aplikacije rudarenja teksta uključuju filtere elektroničke pošte i interesnih grupa.

2.4.5 Prepoznavanje slika

Obrada i klasifikacija digitalnih slika, statičkih i dinamičkih, pruža velike mogućnosti primjene. Korisno je za raspoznavanje pisanih slova, uspoređivanje i identifikacija ljudskih lica, te otkrivanje sumnjivog ponašanja kamerama za video nadzor.

2.4.6 Rudarenje weba

Koristi se za analize zvane clickstreams, koje prate slijed posjećenih stranica i izbore korisnika interneta. Mogu se pokazati korisnima za analizu komercijalnih web stranica, pronalaženje najpopularnije stranice ili procjenu učinkovitosti tečaja internet učenja.

2.4.7 Medicinske dijagnoze

Modeli učenja su neprocjenjiv alat na području medicine, jer omogućuje rano otkrivanje bolesti korištenjem rezultata kliničkih testova.

2.5 Prikaz ulaznih podataka

U većini slučajeva, ulazni podaci za analizu rudarenja podataka imaju oblik dvodimenzionalnih tablica, koje se nazivaju set podataka (*engl. dataset*). Redovi u setu podataka predstavljaju zapise zapažanja iz prošlosti, i nazivaju se primjerima, slučajevima, instancama ili zapisima. Kolone predstavljaju informacije o pojedinom promatranju, i nazivaju se atributima, varijablama, karakteristikama ili značajkama.

Atributi sadržani u setu podataka mogu se svrstati po kategorijama ili numerički, ovisno o vrsti vrijednosti koje poprimaju.

2.5.1 Kategorički

Kategoričke atribute predstavlja konačan broj različitih vrijednosti, koji je obično ograničen na manje od sto, i predstavljaju kvalitativna svojstva entiteta na koje se odnose. Primjerice, atribut županija u kojoj boravi neki pojedinac poprima jednu od vrijednosti iz predodređenog niza, moguće je da ta vrijednost bude cijeli broj. Drugi primjer su telefonske kompanije, prelazak klijenta na mrežu drugog operatera označava se sa vrijednošću 1, dok se ostanak na mreži označava sa 0. Na kategoričke atribute se ne primjenjuju aritmetičke operacije, čak ni kad su kodirane vrijednosti izražene cijelim brojevima.

2.5.2 Numerički

Numeričke atribute predstavljaju konačne ili beskonačne brojčane vrijednosti, nad kojima je moguće obavljati operacije oduzimanja ili dijeljenja. Primjerice, iznos odlaznih telefonskih poziva klijenta, tijekom jednog mjeseca, prikazan je numeričkom varijablom.

Ponekad i preciznija taksonomija atributa može biti korisna.

2.5.3 Točke

Točke su kategorički atributi čija vrijednost može bit točno ili netočno (*engl. true or false*). Ovi atributi mogu biti izraženi koristeći Boolean varijable {*true, false*} ili binarne varijable {0, 1}. Primjerice, klijent banke može ili ne može koristiti bankovnu kreditnu karticu.

2.5.4 Nominalni

Nominalni atributi su kategorički atributi bez nekog prirodnog slijeda, kao što je županija stanovanja.

2.5.5 Redni

Redni atributi su kategorički atributi koji posuđuju svojstvo prirodnog slijeda, ali nema smisla računati razlike ili omjere između vrijednosti. Primjer za ove atribute je stupanj obrazovanja.

2.5.6 Diskretni

Diskretni atributi su numerički atributi koje predstavljaju konačni brojevi ili brojive beskonačne vrijednosti.

2.5.7 Kontinuirani

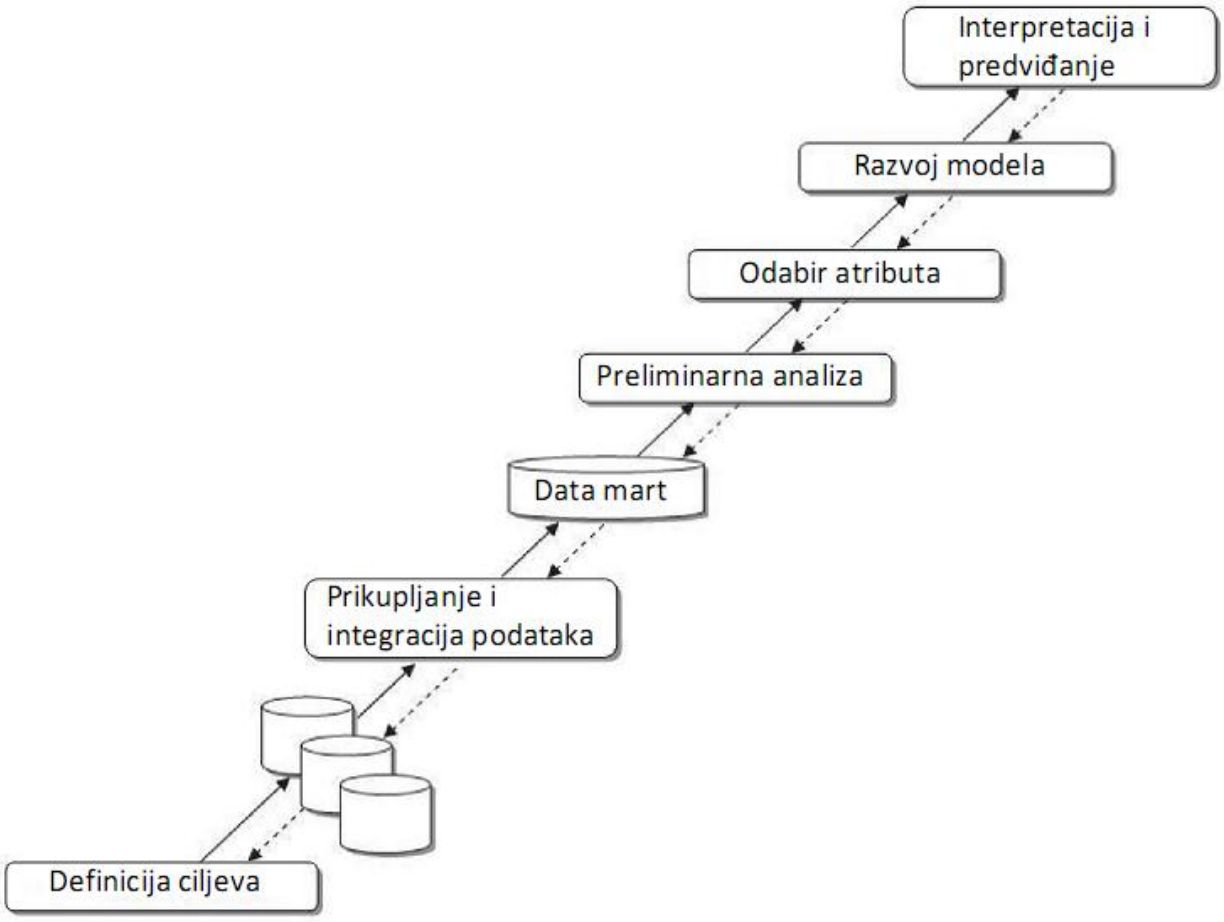
Kontinuirani atributi su numerički atributi koje predstavljaju nebrojive beskonačne vrijednosti.

2.6 Proces rudarenja podataka

Definicija rudarenja podataka koja je dana na početku ovog poglavlja odnosi se na iterativni proces, u kojem modeli učenja i tehnike imaju ključnu ulogu. Slika 5.1 prikazuje glavne faze procesa rudarenja podataka.

2.6.1 Definicija ciljeva

Analize rudarenja podataka provode se određenim stručnim područjima i cilj im je donositeljima odluka pružiti korisno znanje. Zbog toga je potrebna određena sposobnost stručnjaka s tog područja da dobro definira ciljeve istraživanja. Ako problem nije adekvatno identificiran i ograničen, postoji mogućnost da će sve buduće aktivnosti rudarenja podataka biti izvođene uzalud. Ciljevi će biti bolje definirani ako postoji bliska suradnja između stručnjaka na područja gdje se primjenjuje analiza i analitičara rudarenja podataka.



Slika 2-1. Proces rudarenja podataka

2.6.2 Prikupljanje i integracija podataka

Podaci mogu potjecati iz različitih izvora, i zato može biti potreba integracija. Izvori podataka mogu biti unutarnji, vanjski ili kombinacija ta dva tipa. Integracija različitih izvora podataka može biti potrebna onda kada podatke treba obogatiti s novim dimenzijama, kao što su geomarketinške varijable, liste potencijalnih kupaca, koje još ne postoje u informacijskom sustavu tvrtke. U nekim slučajevima, izvori podataka su već strukturirani u skladištima podataka ili data marts,i koriste se za OLAP analize i aktivnosti za podršku donošenja odluka. To su vrlo povoljne situacije gdje je dovoljno odabrati atribute koji su potrebni za analizu

rudarenja podataka. Međutim, moguće je da su ti podaci iz skladišta podataka agregirani i konsolidirani do te mjere da su beskorisni za bilo kakve analize. Primjerice, ako maloprodajno poduzeće pohranjuje samo ukupan iznos računa, bez da bilježi pojedine artikle, moguće je da će analize rudarenja podataka koje su usmjerene na istraživanje kupovine biti ugrožene.

2.6.3 Preliminarna analiza

Preliminarna analiza izvodi se kako bi proučili dostupni podaci i provelo čišćenje podataka. Uglavnom su podaci pohranjeni u skladišta podataka obrađeni za vrijeme učitavanja, na način da su uklonjene sve sintaktičke pogreške. Primjerice, datumi rođenja koji nisu u prihvatljivom razdoblju su uklonjeni, ili negativne naplate u prodaje su ispravljene. U rudarenju podataka čišćenje podataka provodi se na semantičkoj razini. Najprije se proučava raspon vrijednosti svakog atributa, korištenjem histograma za kategoričke atribute i osnovne statistike za numeričke vrijednosti. Na taj način se otkrivaju sve abnormalne vrijednosti i vrijednosti koje nedostaju. Te podatke proučavaju stručnjaci s tog područja kako bi razmotrili isključivanje pojedinog zapisa iz istraživanja.

2.6.4 Odabir atributa

Važnost atributa se ocjenjuje ovisno o ciljevima analize. Atributi koji za koje se pokaže da nisu od neke važnosti su uklonjeni, kako bi se seta podataka uklonile nevažne informacije. Zatim slijedi prikladna transformacija i uključivanje novih atributa u set podataka. Primjerice, u većini slučajeva je korisno uvesti nove atribute koji odražavaju trendove podataka, a dobivaju se izračunavanjem omjera i razlika između originalnih varijabli. Preliminarna analiza i odabir atributa su kritične faze procesa rudarenja podataka i mogu uvelike utjecati na uspjeh narednih faza.

2.6.5 Razvoj modela

Kad se sastavi kvalitetan set podataka, moguće je početi razvoj modela za raspoznavanje obrazaca i predviđanje. Uglavnom se osposobljavanje (*engl. training*) modela provodi koristeći uzorak zapisa izvađenih iz originalnog seta podataka. Zatim se točnost predviđanja svakog modela procjenjuje pomoću ostatka podataka, koji se dijele na dva dijela. Prvi se sastoji od seta za osposobljavanje (*engl training set*) i koristi se za identifikaciju specifičnih modela učenja unutar odabrane klase modela. Obično se odabire relativno mala količina uzoraka seta za osposobljavanje, koja je statistički vrlo značajna (nekoliko tisuća uzoraka). Drugi set podataka je test set i koristi se za procjenu točnosti alternativnih modela

generiranih tijekom faze osposobljavanja, kako bi se identificirao najbolji model buduća predviđanja.

2.6.6 Interpretacija i predviđanje

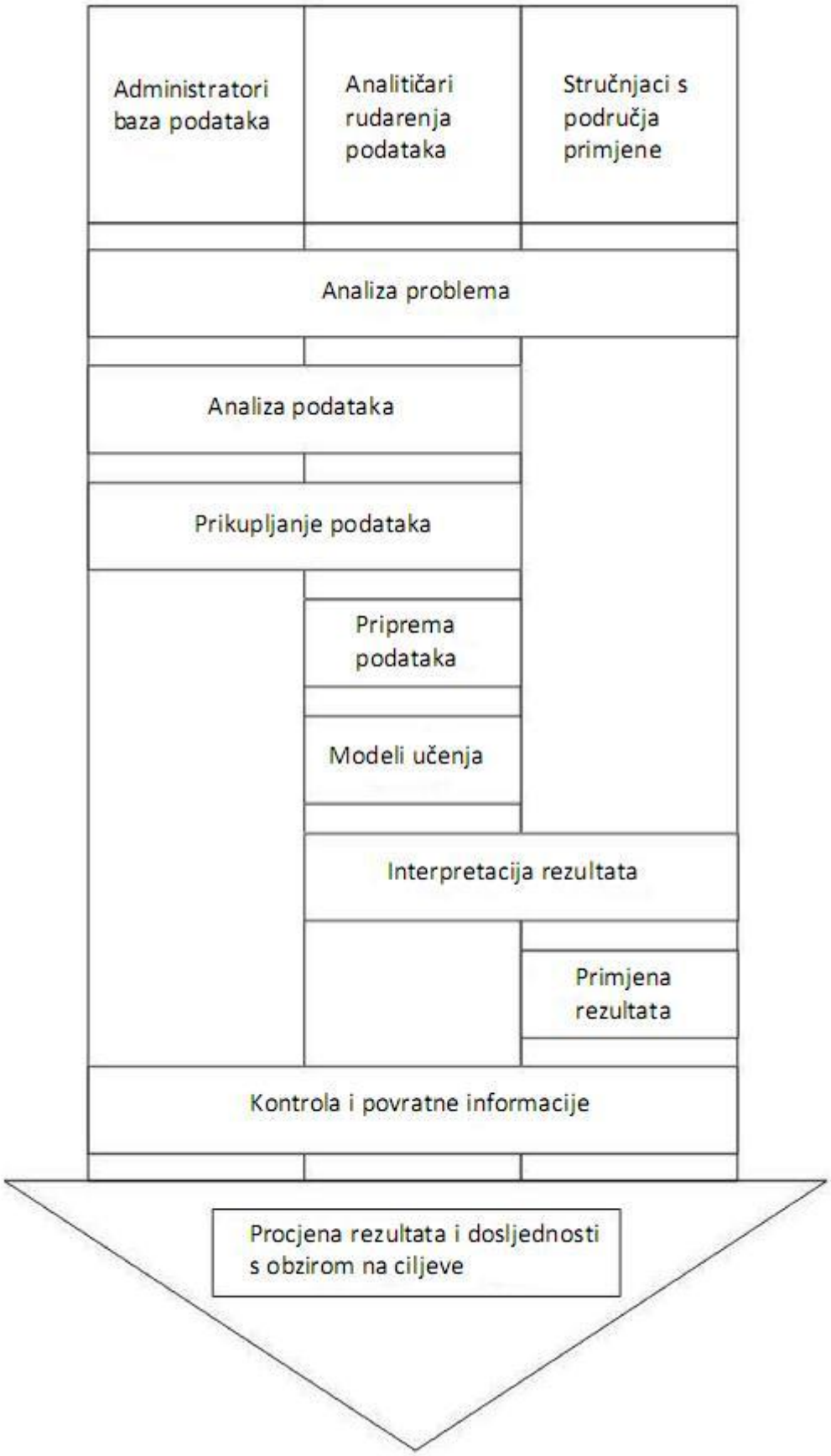
Na kraju procesa rudarenja podataka, koristi se odabrani model, kako bi se postigli zadani ciljevi. Taj se model može primjenjivati i za procedure podrške procesa donošenja odluka, kako bi donositelji odluka imali uvid u predviđanja i stekli veća znanja od proučavanom problemu.

Proces rudarenja podataka uključuje povratne informacije, prikazane isprekidanim linijama na slici 4.1, koje ukazuju na povratak na prethodne faze ovisno o ishodu naredne.

Treba naglasiti važnosti uključivanja i suradnje više različitih stručnjaka, kako bi se proces rudarenja podataka izveo učinkovito:

- stručnjak područja koje se istražuje, od kojeg se očekuje da definira ciljeve analize, da objasni značenje podataka i da sudjeluje u odabiru najučinkovitijeg i najpreciznijeg modela
- stručnjak za informacijski sustav tvrtke, od kojeg se očekuje da nadzire pristup izvorima informacija
- stručnjak za matematičke teorije učenja i statistiku, koji je potreban za provedbu preliminarne analize i generiranje modela za predviđanje

Slika 2-2 prikazuje nadležnosti u različitim aktivnostima za svakog sudionika u procesu rudarenja podataka.



Slika 2-2. Sudionici i njihove uloge u procesu rudarenja podataka

2.7 Analiza metodologija

Aktivnosti rudarenja podataka mogu se podijeliti u nekoliko glavnih kategorija, što zavisi o zadacima i ciljevima analize. Ovisno o postojanju tražene varijable, moguće je podijeliti procese učenja na nadzirane (*engl. supervised*) i nenadzirane (*engl. unsupervised*).

2.7.1 Nadzirani procesi učenja

U nadziranoj (direktnoj) analizi učenja ciljani atribut predstavlja klasu kojoj zapis pripada, ili izražava mjerljivu veličinu, kao što je primjerice ukupna vrijednost poziva koje će klijent ostvariti u budućem razdoblju. Nadzirani procesi učenja su orijentirani predviđanju i interpretaciji s obzirom na određeni atribut.

2.7.2 Nenadzirani procesi učenja

Nenadzirani (indirektni) procesi učenja nisu vođeni ciljanim atributom. Zato je, u ovom slučaju, zadatak rudarenja podataka pronaći ponavljajuće obrasce i sklonosti skupa podataka. Primjerice, investicijska kompanija, na temelju transakcija iz prošlosti, želi identificirati skupinu klijenata koji imaju slična investicijska ponašanja. Većini nenadziranih procesa učenja cilj je identificirati skupine zapisa koji su slični unutar svake skupine i različiti od ostalih skupina.

Postoji sedam osnovnih zadaća rudarenja podataka:

- karakterizacija i diskriminacija
- klasifikacija
- regresija
- analiza vremenskih nizova
- asocijativna pravila
- grupiranje
- opis i vizualizacija

Prva četiri zadatka, zadaci su nadziranih analiza rudarenja podataka, koji se koriste ako postoje određene ciljane varijable koje je potrebno obrazložiti na temelju dostupnih atributa ili tijekom razvoja. Ostala tri su zadaci nenadziranih analiza čija je svrha razviti model sposoban izraziti međuodnose dostupnih atributa.

2.7.3 Karakterizacija i diskriminacija

Tamo gdje postoji specifični ciljani atribut, često se prije razvoja klasifikacijskog modela provodi preliminarna analiza, i to zbog dva razloga. Prvi razlog tome je karakterizacija atributa zapisa koji pripadaju istoj klasi, uspoređujući raspodjelu njihovih vrijednosti. Drugi razlog je otkrivanje razlike, uspoređujući raspodjelu vrijednosti atributa

zapisa između zadane klase i zapisa drugačije klase, ili između zapisa zadane klase i preostalih zapisa. Ovaj zadatak rudarenja podataka provodi se putem preliminarne analize podataka, i stoga se temelji na upitima koji ne zahtijevaju razvoj specifičnih modela učenja. Informacije koje su prikupljene na ovaj način, korisnicima se prikazuju pomoću histograma i drugih vrsta grafova. Vrijednost generiranih informacija je izvanredna i često se određuje naknadna faza odabira atributa.

2.7.4 Klasifikacija

Kod ovog problema dostupan je skup zapažanja, uglavnom taj skup predstavljaju zapisi seta podataka, čija je ciljana klasa poznata. Primjerice, zapažanja telefonskih kompanija mogu biti binarne klase koje ukazuju da li je određeni klijent aktivan ili je prešao na mrežu konkurencije. Svako zapažanje opisano je određenim brojem atributa čije su vrijednosti poznate. U navedenom primjeru, atributi mogu biti dob, vrijeme koje kupac koristi uslugu i odlazni pozivi. Dakle, algoritmi za klasifikaciju koriste dostupna zapažanja kako bi identificirali model koji može predvidjeti buduću ciljanu klasu čije su vrijednosti atributa poznate. Važno je napomenuti da je ciljani atribut čija se vrijednost predviđa kategoričan u klasifikacijskim problemima i stoga poprima konačan i vrlo mali broj vrijednosti. U većini slučajeva su vrijednosti ciljanih atributa prikazane binarnim varijablama. Kategorička narav ciljanog atributa razlikuje klasifikaciju od regresije.

2.7.5 Regresija

Za razliko od klasifikacije, koja je namijenjena diskretnim vrijednostima, regresija se koristi kada ciljana varijabla poprima kontinuirane vrijednosti. Ovisno o dostupnosti atributa, cilj je predvidjeti vrijednost ciljane varijable za svako zapažanje. Primjerice, ako se želi predvidjeti prodaja proizvoda na temelju promotivnih kampanja i prodajnoj cijeni, ciljana varijabla može poprimiti vrlo visok broj diskretnih vrijednosti te se može tretirati kao kontinuirana varijabla. Klasifikacijski problem se može pretvoriti u regresijski, i obrnuto.

2.7.6 Vremenski nizovi

Ponekad ciljani atribut tijekom vremena evoluira i postane povezan s susjednim vremenskim periodima. U tom slučaju redoslijed vrijednosti ciljane varijable predstavlja vremenski niz. Primjerice, dvije godine se promatra tjedna prodaja određenog proizvoda i predstavlja ju vremenski niz koji sadrži 104 zapažanja. Model za analizu vremenskog niza

ispituje podatke, koje karakterizira tijek vremena, i usmjeren je na predviđanje vrijednosti ciljane varijable za jedno ili više budućih razdoblja.

2.7.7 Asocijativna pravila

Asocijativna pravila (grupacije) se koriste za identifikaciju zanimljivih i ponavljajućih asocijacija između grupa zapisa u setu podataka. Primjerice, moguće je odrediti koji se proizvodi kupuju zajedno i koliko često se to događa. Maloprodajne kompanije koriste asocijativna pravila za slaganje proizvoda na policama ili za uređenje kataloga. Grupacije povezanih elemenata također se koriste za kombiniranje proizvoda i usluga.

2.7.8 Grupiranje

Tehnike grupiranja usmjerene su na segmentiranje heterogene populacije u određeni broj podgrupa sastavljenih od opažanja koja dijele slične karakteristike. Za razliku od klasifikacije, grupiranje nema predefinirane klase ili primjere koji ukazuju na ciljanu klasu, tako da su objekti grupirani ovisno o međusobnoj homogenosti. Ponekad identificirane grupe predstavljaju preliminarnu fazu procesa rudarenja podataka, koja se nalazi unutar preliminarne analize podataka. To omogućuje obradu homogenih podataka odgovarajućim tehnikama, te je također moguće smanjiti set originalnih podataka, kako bi se naknadne aktivnosti rudarenja podataka mogle razvijati neovisno za svaku grupu.

2.7.9 Opis i vizualizacija

Ponekad je cilj rudarenja podataka pružiti jednostavan i sažet prikaz podataka pohranjenih u velikom skupu podataka. Iako za razliku od grupiranja i asocijativnih pravila ovaj zadatak ne teži nekoj određenoj grupaciji zapisa u setu podataka, učinkovit i sažet opis informacija vrlo je koristan, jer može ukazati na objašnjena nekih obrazaca i dovesti do boljeg razumijevanja zadanog problema. No nije uvijek lako dobiti smislenu vizualizaciju podataka. Međutim, trud koji je uložen u vizualni prikaz podataka opravdavaju izvanredne informacije koje su dobivene pomoću dobro osmišljenog grafa.

3. Rudarenje podataka pomoću Business Intelligence Development Studio-a

Jedan od najizazovnijih aspekata rudarenja podataka pomoću SSAS-a je razumjeti što je moguće postići pomoću različitih algoritama, a zatim kreirati strukturu rudarenja podataka koja obuhvaća prikladan odnosno prikladne algoritme koji najbolje odgovaraju zahtjevima poslovanja. Između ostalog, bitan je i način na koji su dobiveni podaci prikazani krajnjim korisnicima. [1]

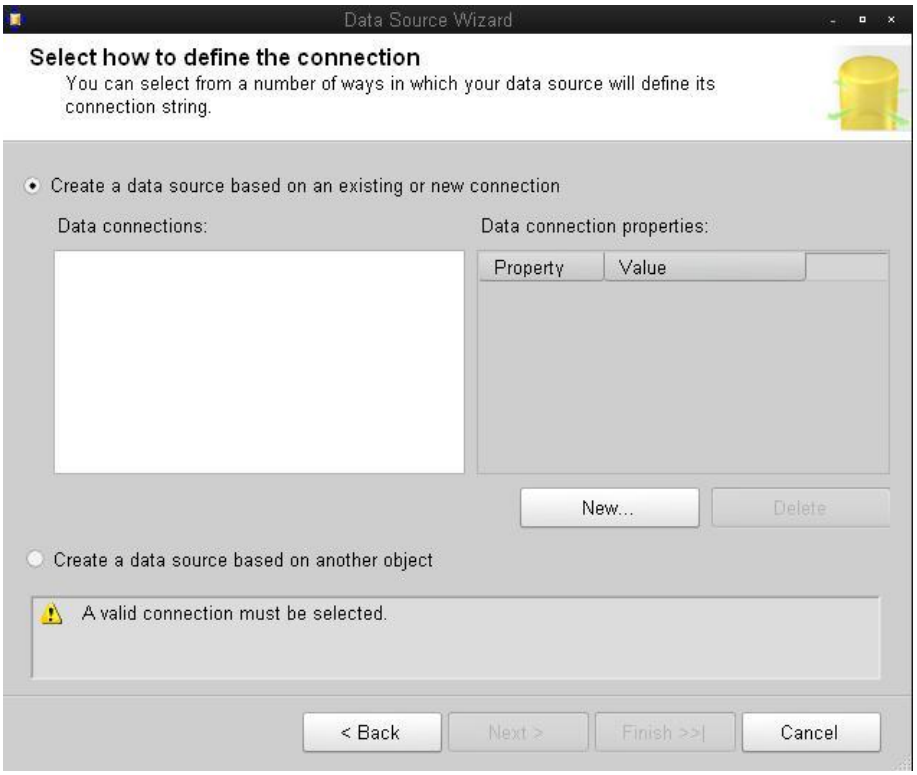
3.1 Sučelje BIDS-a

Za demonstraciju rudarenja podataka pomoću BIDS-a korišten je AdventureWorks BI primjer. Primjer uključuje pet struktura rudarenja podataka. Svaka struktura uključuje jedan ili više modela rudarenja podataka, i svaki model temeljen je na jednom od Microsoft-ovih algoritama rudarenja podataka.

Prije izrade strukture rudarenja potrebno je definirati Data Source i Data Source View. Unutar Solution Explorera desnom klikom miša iznad Data Sources odabire se opcija New Data Source, kao što je prikazano na slici 3-1.

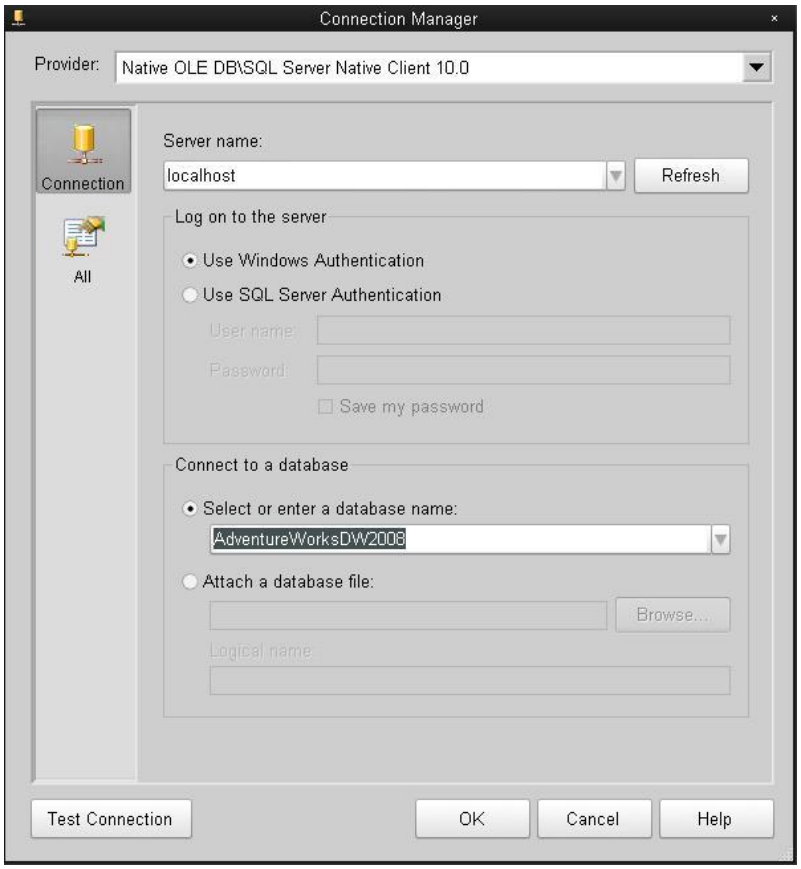
Slika 3-1. Definiranje izvora podataka-Data Source

Odabirom opcije New Data Source pokreće se Data Source Wizard, prikazano na slici 3-2, nakon čega je potrebno definirati ime servera i baze podataka u kojima se nalaze podaci potrebni za daljnje analize, slika 3-3.

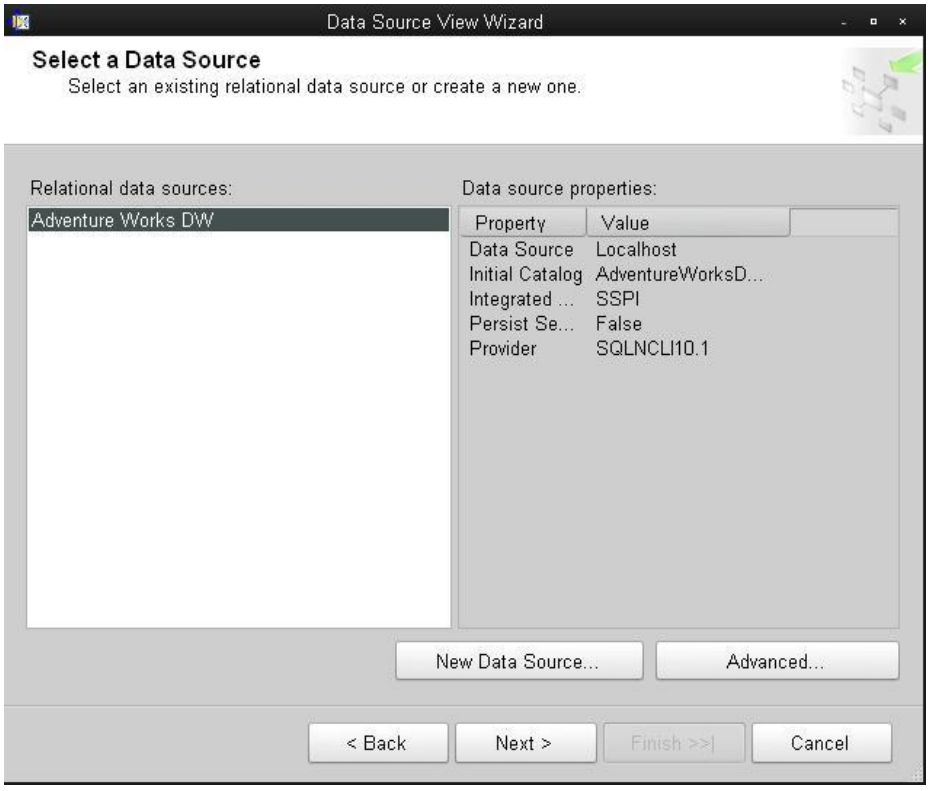


Slika 3-2. Data Source Wizard

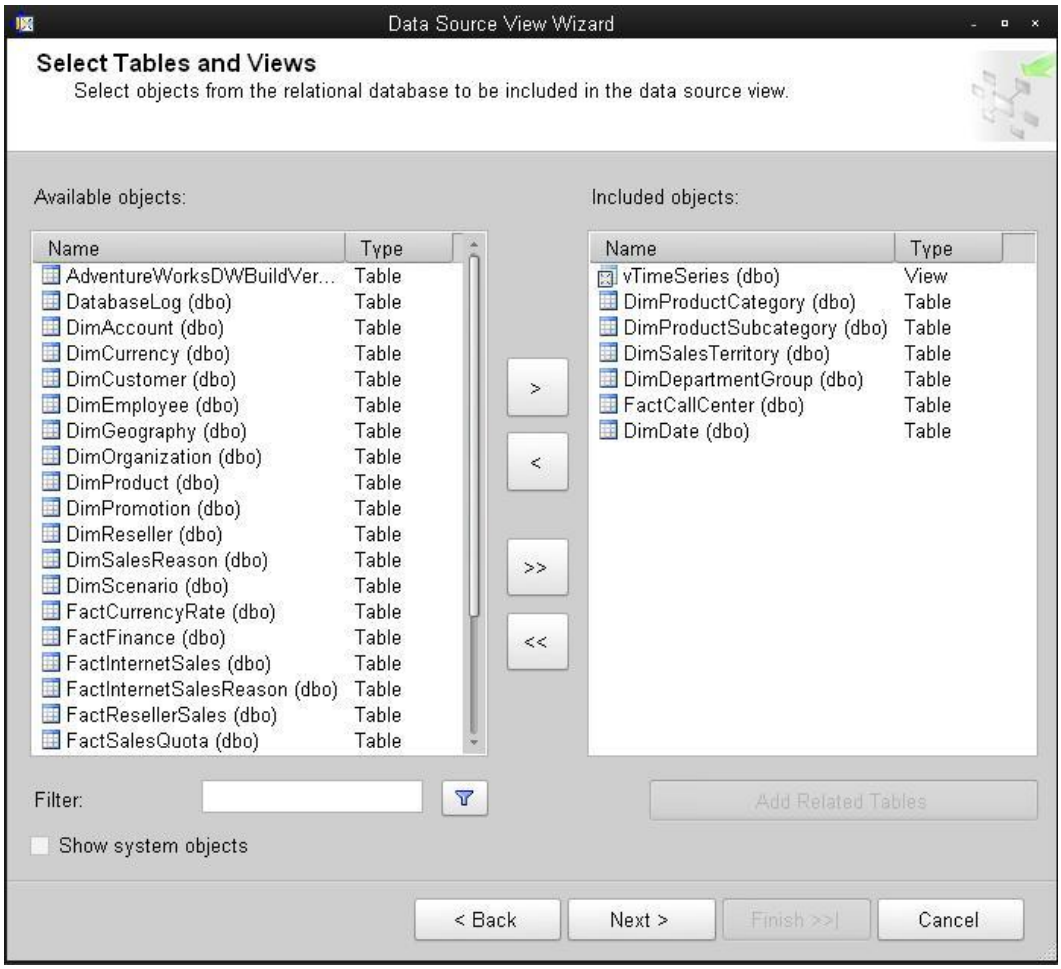
Kada je definiran Data Source, preostaje još definiranje Data Source View-a. Data Source View Vizard pokreće se na isti način kao i Data Source, odabirom opcije New Data Source View nakon desnog klika na Data Source Views unutar Solution Explorer-a. Na slici 3-4 prikazan je Data Source View Wizard u kojem se nalazi lista ponuđenih (prethodno definiranih Data Source-a). U ovom koraku moguće je dodati novi izvor podataka ukoliko je to potrebno. U sljedećem koraku, slika 3-5, odabiru se objekti iz prethodno odabrane baze podataka s čime završava definiranje Data Source View-a.



Slika 3-3. Definiranje servera i baze podataka

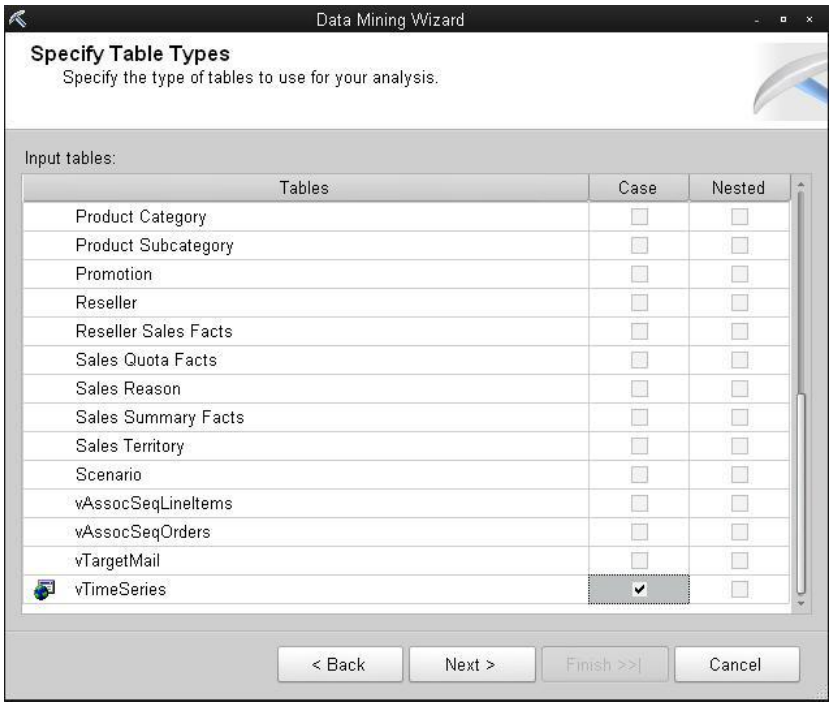


Slika 3-4. Data Source View Wizard



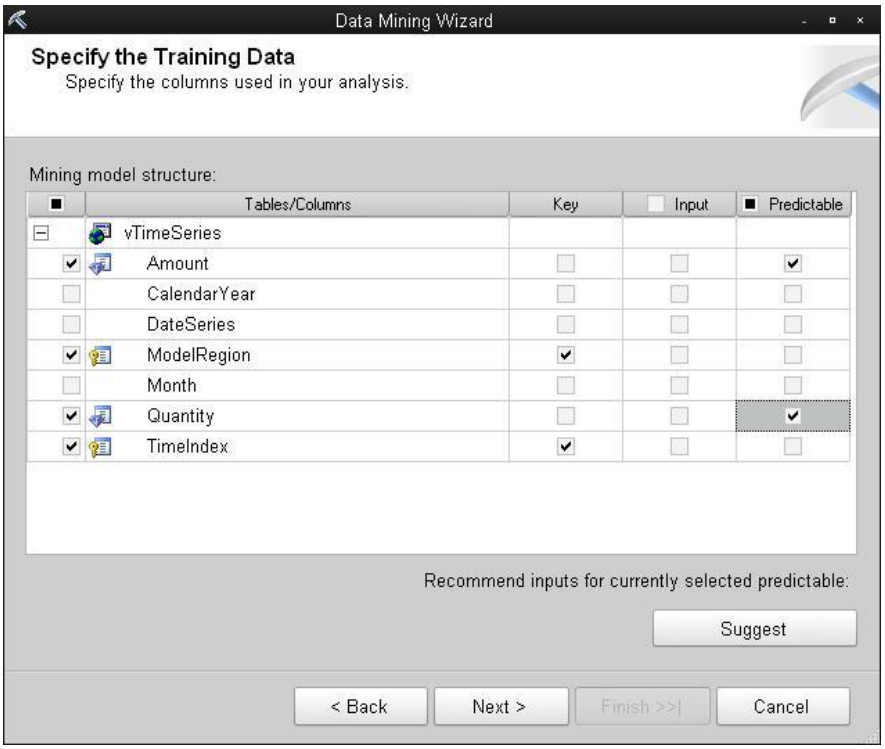
Slika 3-5. Data Source View Wizard - odabir objekata

Nakon što su definirani Data Source i Data Source View moguće je pristupiti izradi strukture modela rudarenja podataka. Novu strukturu modela rudarenja podataka također je moguće dodati pomoću čarobnjaka, odnosno Data Mining Wizard-a. U prvom koraku je potrebno odabrati oblik izvora podataka između postojeće relacijske baze podataka ili skladišta podataka, i postojeće kocke. Sljedeći korak je odabir algoritma rudarenja podataka (detaljan opis pojedinog algoritma u poglavlju 3.4), odnosno odabir izrade strukture bez algoritma (koje je moguće kasnije dodati). Treći korak je odabir između ponuđenih Data Source View-a, nakon čega je potrebno odabrati tablice nad čijim će se podacima vršiti rudarenje podataka. Dijalog za odabir tih tablica prikazan je na slici 3-6.

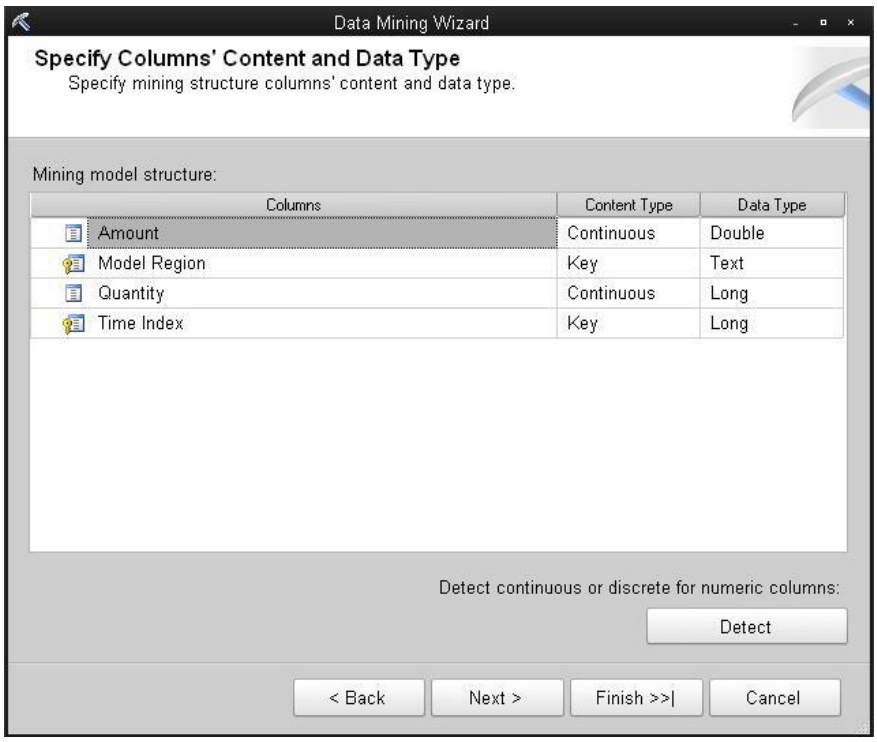


Slika 3-6. Odabir tablica za rudarenje podataka

U sljedećem koraku potrebno je definirati ulazne kolone, odnosno parametre, te vrijednosti koje želimo predvidjeti odnosno izračunati, prikazano na slici 3-7. Slika 3-8 prikazuje korak u kojem se definiraju vrste podataka i vrste sadržaja (o tome više u sljedećem poglavlju). Detekciju vrsta podataka i vrsta sadržaja moguće je prepustiti softveru.



Slika 3-7. Definiranje kolona



Slika 3-8. Definiranje vrste sadržaja i vrste podataka

Nakon toga preostaje dodjela naziva strukturi što je posljednji korak Data Mining Wizarda.

3.2 Vrste podataka i vrste sadržaja

Analysis Services strukture rudarenja podataka koriste specifične vrste podataka i sadržaja za implementaciju rudarenja podataka. Za izradu struktura rudarenja potrebno je razumijevanje tih vrsta. Također, neki algoritmi podržavaju samo neke vrste sadržaja.

Moguće vrijednosti vrste podataka su Text, Long, Boolean, Double, ili Date. Vrste podataka su detektirane i dodijeljene automatski prilikom izrade strukture rudarenja podataka.

Vrsta sadržaja je dodatan atribut koji koristi algoritam rudarenja podataka, kako bi pobliže shvatio ponašanje podataka. Primjerice, ako je sadržaj izvorne kolone označen kao Cyclical, algoritam rudarenja zna da se radi o specifičnim podacima koji se ponavljaju. Jedan primjer toga je broj mjeseci za više od jedne godine unutar vremenske tablice.

Najprije se određuje vrsta podataka, a zatim prikladan tip sadržaja modela. Valja zapamtiti da određeni algoritmi podupiru određene vrste sadržaja. Primjerice, Naïve Bayes algoritam ne podupire Continious vrstu sadržaja. Data Mining Wizard detektira vrstu sadržaja prilikom

izrade strukture rudarenja. Slijedi opis vrsta sadržaja i popis vrsta podataka koje je moguće koristiti s određenim vrstama sadržaja.

Discrete

Ovako označena kolona sadrži različite vrijednosti, primjerice, određeni broj djece. Ne sadrži nepotpune vrijednosti. Također, Discrete kolona ne ukazuje na važnost redoslijeda (ili niza) informacija. S ovom vrstom sadržaja moguće je koristiti sve vrste podataka.

Continious

Ovako označena kolona sadrži brojeve čije vrijednosti predstavljaju neku jedinicu mjere. Te vrijednosti mogu biti djelomične. Primjerice, preostali iznos kredita. S ovom vrstom sadržaja moguće je koristiti Date, Double, ili Long vrstu podataka.

Discretized

Ovako označena kolona sadrži kontinuirane vrijednosti koje su grupirane u ćelije. Za svaku se ćeliju smatra da ima određeni redoslijed i da sadrži diskretne vrijednosti. Na slici 8.9 vidljivo je kako je za Targeted Mining primjer korištena kolona Age. Za kolonu je moguće definirati parametar DiscretizationMethod, i DiscretizationBucketCount (ako je vrsta sadržaja

definirana kao Discretized). U primjeru na slici je za parametar DiscretizationBucketCount zadana vrijednost 10, a za DiscretizationMethod vrijednost Automatic. Moguće vrijednosti DiscretizationMethod parametra su Automatic, Equal Areas, ili Clusters. Automatic znači da SSAS određuje koja će se metoda koristiti. Equal Areas znači da će ulazni podaci biti podijeljeni na dva jednaka dijela. Ova metoda najbolje funkcionira s pravilno raspoređenim podacima. Clusters znači da SSAS uzorkuje podatke kako bi rezultat bio nakupina vrijednosti podataka. Zbog uzorkovanja ova metoda koristi samo numeričke ulaze. S ovom vrstom sadržaja moguće je koristiti Date, Double, Long, ili Text vrstu podataka.

Table

Ova vrsta kolone sadrži ugniježđenu tablicu koja ima jednu ili više kolona, i jedan ili više redova. Te kolone mogu sadržavati više vrijednosti, ali jedna od tih vrijednosti mora biti povezana s nadređenim zapisom. Primjerice, kada je informacija o kupcu u tablici slučaja povezana s informacijom o kupnji predmeta u ugniježđenoj transakcijskoj tablici.

Key

Ovako označena kolona se koristi kao jedinstveni identifikator reda. S ovom vrstom sadržaja moguće je koristiti Date, Double, Long, ili Text vrstu podataka.

Key Sequence

Ovako označena kolona je vrsta ključa, gdje je redoslijed vrijednosti ključa bitan za model rudarenja. S ovom vrstom sadržaja moguće je koristiti Date, Double, Long, ili Text vrstu podataka.

Key Time

Ovako označena kolona slična je Key Sequence koloni, s time da Key Time ukazuje na to da su vrijednosti povezane s vremenskom skalom. S ovom vrstom sadržaja moguće je koristiti Date, Double, Long, ili Text vrstu podataka.

Ordered

Ovako označena kolona sadrži podatke određenog redoslijeda koji je važan za model rudarenja. Također, SSAS smatra sve podatke diskretnima. S ovom vrstom sadržaja moguće je koristiti sve vrste podataka.

Cyclical

Ovako označena kolona sadrži podatke koji se ponavljaju. To se često koristi s vremenskim vrijednostima (primjerice, mjesec u godini). Ti se podaci smatraju poredanima i diskretnima. S ovom vrstom sadržaja moguće je koristiti sve vrste podataka.

Tablica 3-1 sadrži listu tipova podataka i sve vrste sadržaja koje podupiru. Definiranje odgovarajuće vrste sadržaja i vrste podataka kritična je stvar za izradu modela rudarenja podataka.

Tablica 3-1. Vrste podataka i vrste sadržaja

Vrsta podataka	Moguće vrste sadržaja
Text	Discrete, Discretized, Sequence
Long	Continious, Cyclical, Discrete, Discretized, Key Sequence, Key Time, Ordered (redoslijedom ili vremenom)
Boolean	Discrete
Double	Continious, Cyclical, Discrete, Discretized, Key Sequence, Key Time, Ordered (redoslijedom ili vremenom)
Date	Continious, Cyclical, Discrete, Discretized, Key Time

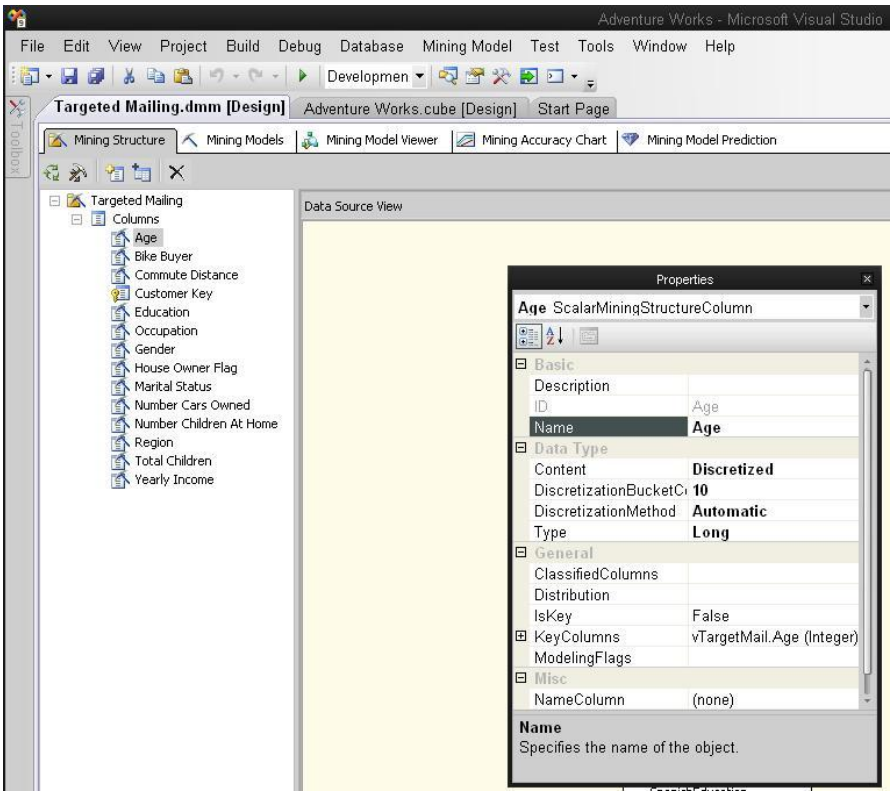
Vrstu podataka i vrstu sadržaja moguće je odrediti pomoću Data Mining Wizard-a ili konfiguracijom u dijalogu Properties.

3.3 Mining Structure tab

Strukturu rudarenja podataka moguće je naknadno mijenjati i prilagođavati. Slika 3-9 prikazuje prvi tab za dizajniranje strukture rudarenja, nazvan Mining Structure. Tu je moguće vidjeti uključeni izvor podataka (Data Source View, DSV) koji je na raspolaganju modelima rudarenja. Međutim, ovdje nisu moguće nikakve promjene DSV-a (promjena naziva kolona, dodavanje izračunatih kolona, itd.). Da bi se napravile bilo kakve promjene DSV-a potrebno je koristiti originalni DSV prikaz. U ovom je pogledu dozvoljeno jedino pregledavati izvor podataka, dodavati ili uklanjati kolone, ili dodavati ugniježđene (engl. nested) tablice strukturi rudarenja.

Moguće je koristiti bilo relacijske tablice ili višedimenzionalne kocke kao izvor podataka za strukturu rudarenja podataka. Ako se odaberu relacijske tablice, ti se podaci mogu preuzeti iz jedne ili više relacijskih tablica, svaku pomoću primarnog ključa. Također je moguće kao izvor podataka odabrati ugniježđenu tablicu. Primjer za to su tablice kupaca i njihovih narudžbi. Tablica Customers je tablica slučajeva, a tablica Orders ugniježđena tablica. Ova situacija zahtjeva primary key/foregin key vezu između redova tih tablica. Jedan

način za dodavanje ugniježđene tablice u strukturu rudarenje je da se nakon desnog klika mišom u Object Browser-u odabere opcija Add a Nested Table.



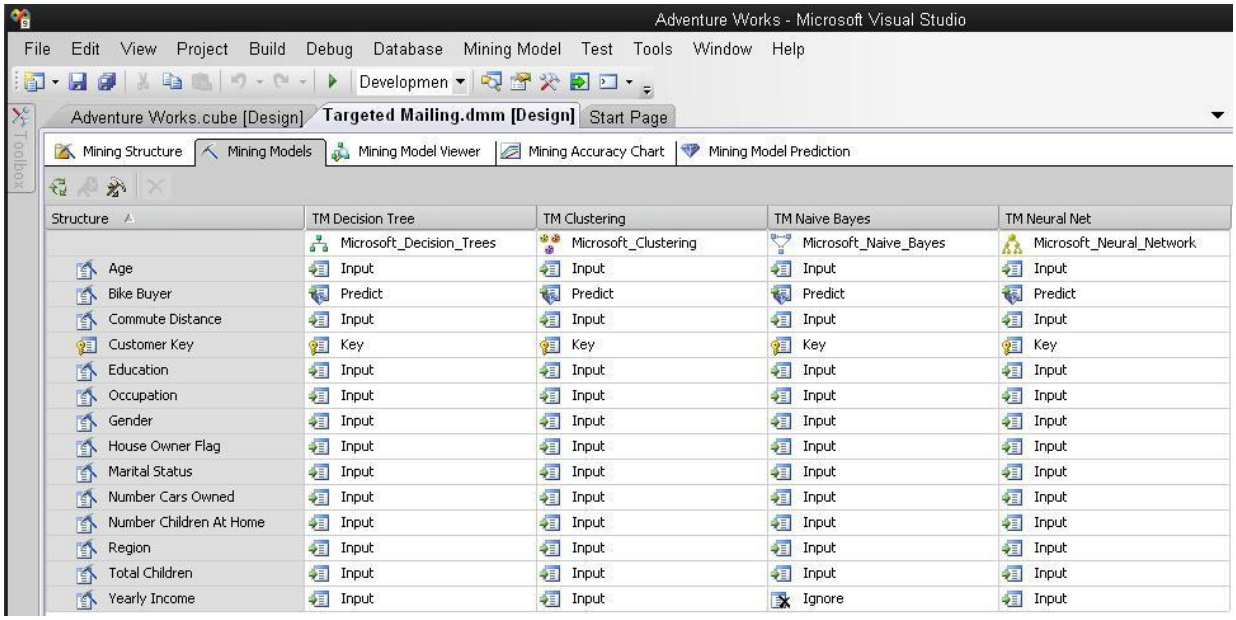
Slika 3-9. BIDS sučelje za strukture rudarenja podataka

U ovom pogledu moguće je konfigurirati nekoliko svojstva strukture rudarenja podataka. Primjerice svojstvo CacheMode. Moguće je odabrati između dvije opcije KeepTrainingCases ili ClearAfterProcessing. Potonja se često koristi tijekom rane faze razvoja projekta. Postoji i mogućnost procesuiranja individualnog modela rudarenja samo da se otkrije trebaju li podaci dodatno čišćenje. U tom slučaju se obavlja naknadno čišćenje, i zatim ponovno procesuiranje modela. Alternativno tome, moguće je provesti potpuni postupak procesuiranja nad cijelom strukturom rudarenja. Po završetku toga, svi modeli rudarenja koji su definirani unutar odabrane strukture su procesuirani.

3.4 Mining Models tab

Sljedeći tab dizajniranja strukture rudarenja u BIDS-u je Mining Models tab. Ovdje se vide modeli rudarenja koji su uključeni u odabranoj strukturi rudarenja. Novi modeli se jednostavno dodaju u strukturu odabirom opcije New Mining Model nakon desnog klika miša

na radnu površinu. Također je moguće promijeniti izvor podataka povezan s tipom modela rudarenja kreiranjem nove instance istog modela, i/ili „ignoriranjem“ jedne ili više kolona iz strukture rudarenja (DSV-a). Primjer ignoriranja kolone Yearly Income za slučaj Microsoft Naïve Bayes algoritma prikazan je na slici 3-10.



Structure	TM Decision Tree	TM Clustering	TM Naive Bayes	TM Neural Net
	Microsoft_Decision_Trees	Microsoft_Clustering	Microsoft_Naive_Bayes	Microsoft_Neural_Network
Age	Input	Input	Input	Input
Bike Buyer	Predict	Predict	Predict	Predict
Commute Distance	Input	Input	Input	Input
Customer Key	Key	Key	Key	Key
Education	Input	Input	Input	Input
Occupation	Input	Input	Input	Input
Gender	Input	Input	Input	Input
House Owner Flag	Input	Input	Input	Input
Marital Status	Input	Input	Input	Input
Number Cars Owned	Input	Input	Input	Input
Number Children At Home	Input	Input	Input	Input
Region	Input	Input	Input	Input
Total Children	Input	Input	Input	Input
Yearly Income	Input	Input	Ignore	Input

Slika 3-10. Prikaz uloga kolona

Moguće je promijeniti način korištenja pridruženih (non-key) kolona:

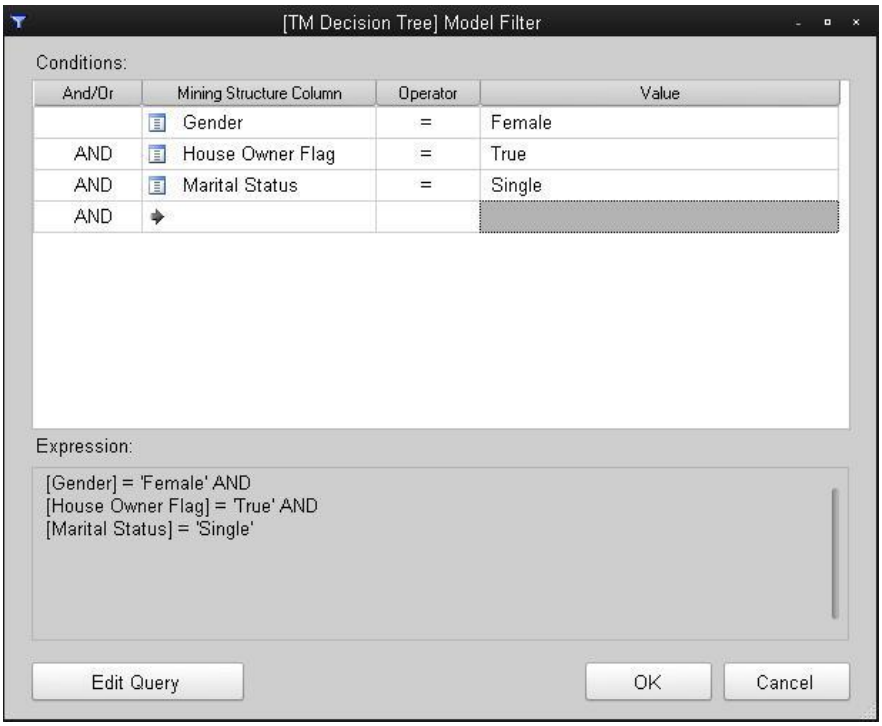
- Ignore – Ova postavka uklanja kolonu iz modela.
- Input – Ova postavka definira kolonu kao izvor podataka za model.
- Predict – Ova postavka definira kolonu kao ulaz i izlaz modela.
- PredictOnly – Ova postavka definira kolonu kao izlaz modela.

Kad su u pitanju ugniježđene tablice treba posebno paziti kada se definira uporaba kolona za algoritme rudarenja podataka. Ako izvor sadrži ugniježđene tablice i ako su označene kao Predict ili PredictOnly, svi ugniježđeni atributi su automatski označeni kao predvidljivi. Zbog toga bi u ugniježđene tablice trebalo uključiti što manji broj atributa.

Mining Models tab je dizajniran da omogući primjenu više modela na iste izvore podataka. Možda u ovom dijelu teksta još i nije jasan razlog zašto je to bitno, jer još algoritmi

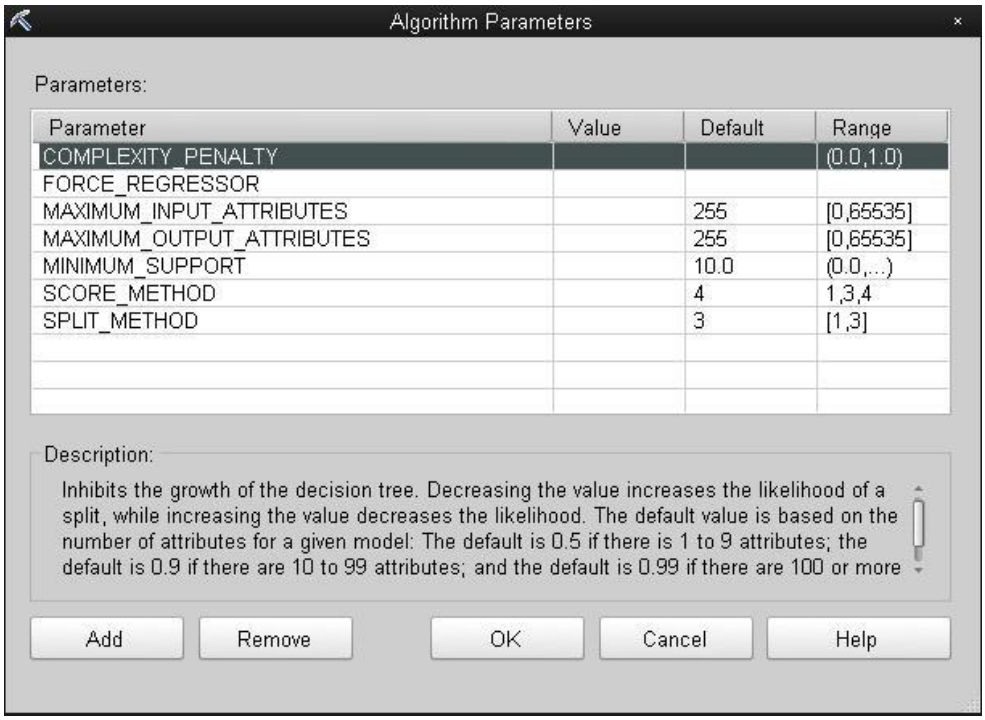
nisu detaljno obrađeni. Potrebno je reći da ova mogućnost omogućuje dodavanje, podešavanje, i uklanjanje modela rudarenja. Analize za predviđanje su više umjetnost nego egzaktna znanost. Primjenjuju se algoritmi za koje se misli da će dati najbolje rezultate. Na početku projekta treba biti spreman na izvođenje velikog proja iteracija kako bi se ispravno podesili svi faktori. Uglavnom su to promjene ulaznih kolona, algoritama, parametara algoritama, itd., dok se ne dođe do korisnih rezultata. SSAS sadrži alate za testiranje i ocjenjivanje svakog modela, kako bi se shvatila korisnost rezultata različitih modela rudarenja.

Nadalje, moguće je kreirati model rudarenja na filtriranom setu izvornih podataka, bez je kreirano više struktura rudarenja. Za implementaciju filtara potrebno je odabrati opciju Set Model Filtar nakon desnog klika na ime modela u BIDS Mining Model tabu. Zatim se pojavljuje dijalog, prikazan na slici 3-11, u kojem se konfiguriraju karakteristike filtara.



Slika 3-11. Konfiguracija filtera

Ugrađena je i mogućnost podešavanja parametara pojedinog modela rudarenja odabirom opcije Set Algorithm Parameters nakon desnog klina na ime modela u BIDS Mining Model tabu. Parametri ovise o odabranom algoritmu rudarenja. Slika 3-12 prikazuje dijalog konfiguracije parametara za Microsoft Decision Trees model.



Slika 3-12. Konfiguracija parametara (MDT model)

Napredni korisnici mogu dodavati vlastite parametre u algoritam putem ovog dijaloga. Ti se parametri razlikuju za svaki uključeni algoritam. Uglavnom promjene tih vrijednosti nisu potrebne, a ako se donose bez razumijevanja mogu rezultirati manjom učinkovitošću odabranog algoritma.

3.5 Mining Model Viewer tab

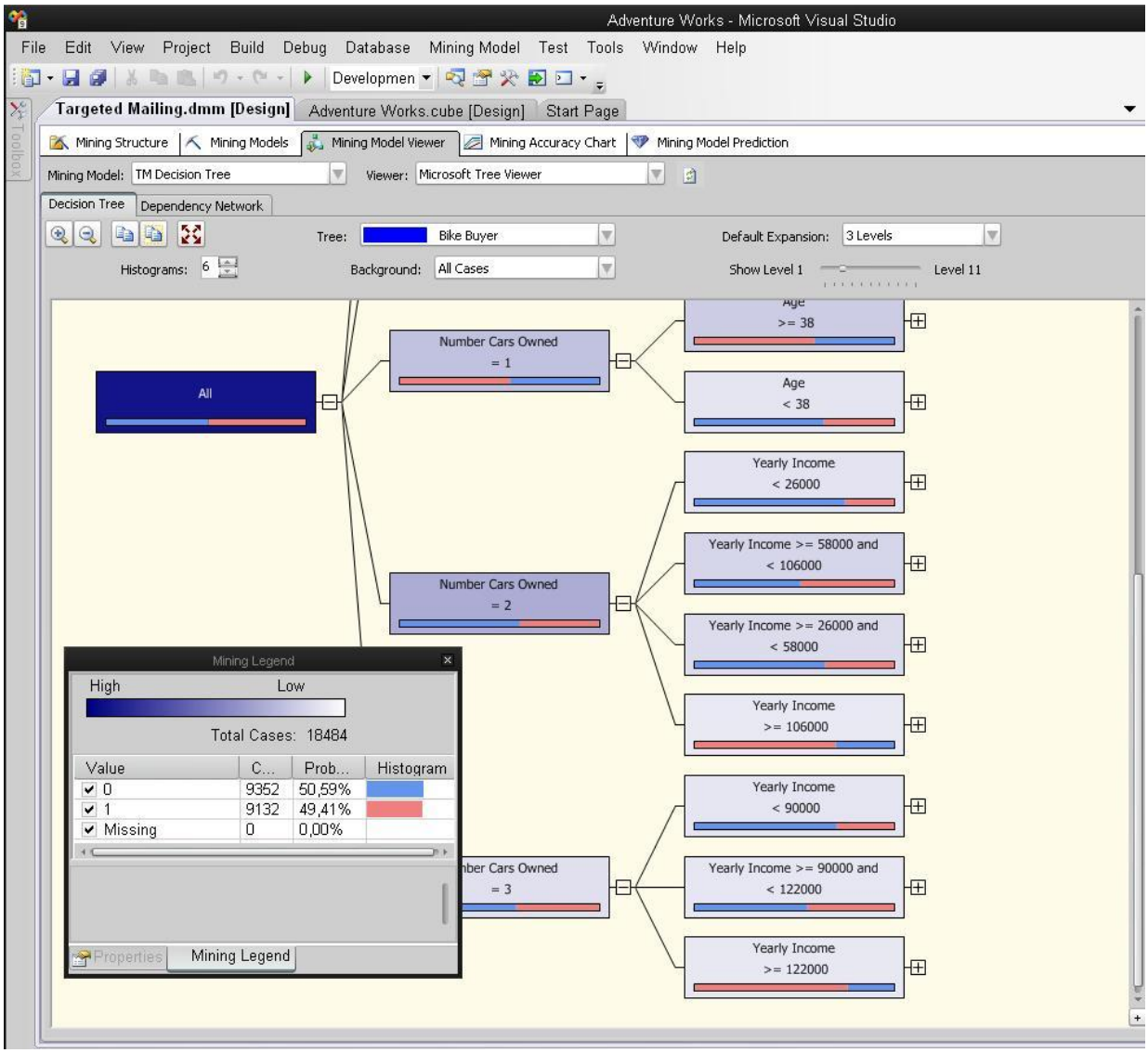
Sljedeći tab u BIDS dizajneru strukture rudarenja je Mining Model Viewer. Zanimljivo je to što svaki algoritam rudarenja uključuje jedan ili više različitih prikaza modela rudarenja. Svrha širokog spektra prikaza pomoći je korisniku da odluči koji su algoritmi rudarenja najkorisniji za njegov poslovni scenarij. Prikazi uključuju grafički i tekstualni (redci i kolone) prikaz podataka. Neki prikazi sadrže više različitih grafičkih prikaza podataka odabranog modela rudarenja. Osim toga, neki sadrže i legendu koja je prikazana u Properties prozoru na radnoj površini. Svaki algoritam ima skup prikaza koji su za njega specifični. Ti prikazi obično predstavljaju informacije pomoću grafova i dijagrama. Osim toga, Microsoft Generic Content Tree Viewer dostupan je svakom algoritmu, i pruža detaljne informacije o svakom čvoru odabranog modela rudarenja.

Slijedi lista algoritama rudarenja podataka dostupnih u SQL Server 2008. Ova lista je samo pregled, te će svaki algoritam detaljno biti prikazan u poglavlju 3.8.

- Microsoft Association
- Microsoft Clustering
- Microsoft Decision Trees
- Microsoft Linear Regression
- Microsoft Logistic Regression
- Microsoft Naïve Bayes
- Microsoft Neural Network
- Microsoft Sequence Clustering
- Microsoft Time Series

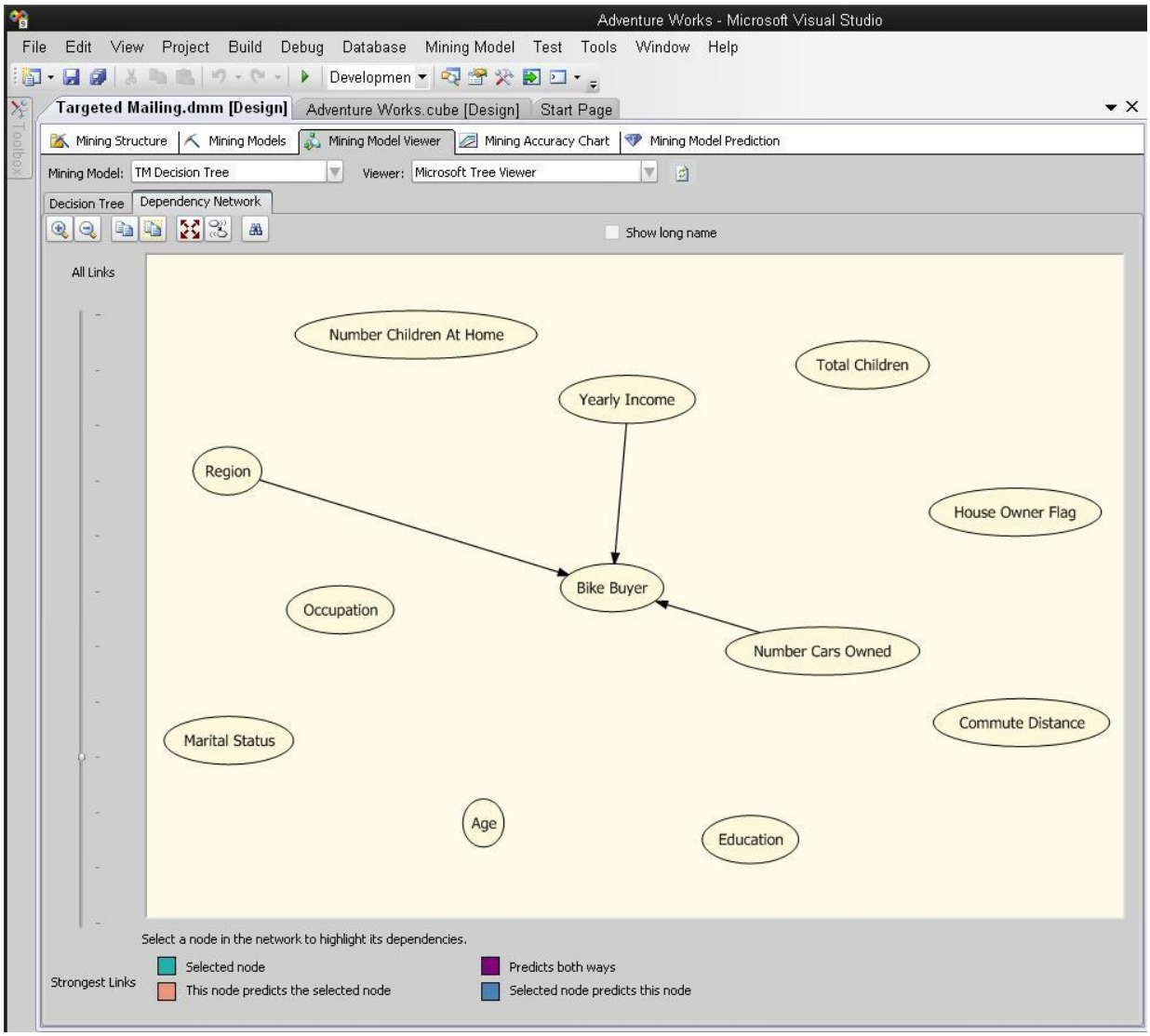
Pomoću AdventureWorks DW 2008 primjera, prikazani su prikazi uključenih algoritama. U primjeru strukture rudarenja zvane Target Mailing moguće je vidjeti četiri različita prikaza jer struktura sadrži četiri modela rudarenja, gdje je svaki model baziran na različitom algoritmu rudarenja. Nakon otvaranja strukture pomoću BIDS-a i otvaranja Mining Model Viewer taba otvatvara se zadani prikaz prvog modela rudarenja koji je na popisu, a to je u ovom slučaju Microsoft Decision Trees algoritam.

Uz prikaz Microsoft Decision Trees algoritma prikazanog na slici 3-13, moguće je dalje prilagođavati prikaz prilagodbom parametara. Slika prikazuje dio Decision Tree pogleda i pripadajuću legendu. Na prvoj razini pokazuje najuže povezane informacije, a u ovom slučaju je to broj automobila koje kupac posjeduje. Dubina boje svakog čvora vizualni je pokazatelj povezanosti, odnosno što je boja tamnija to je veća povezanost. Legenda pokazuje točan broj slučajeva (redaka) za pojedini čvor koji se odabere. Također prikazuje informacije o vjerojatnosti dane u kolonama (postoci) i grafički prikaz istog u sljedećoj koloni. Zadane postavke nivoa (Show Level) su 3. Prikazani model sadrži šest nivoa, međutim prikaz svih bi bilo problematično prikazati na ovoj površini.



Slika 3-13. Microsoft Tree View-er za Microsoft Decision Tree algoritam

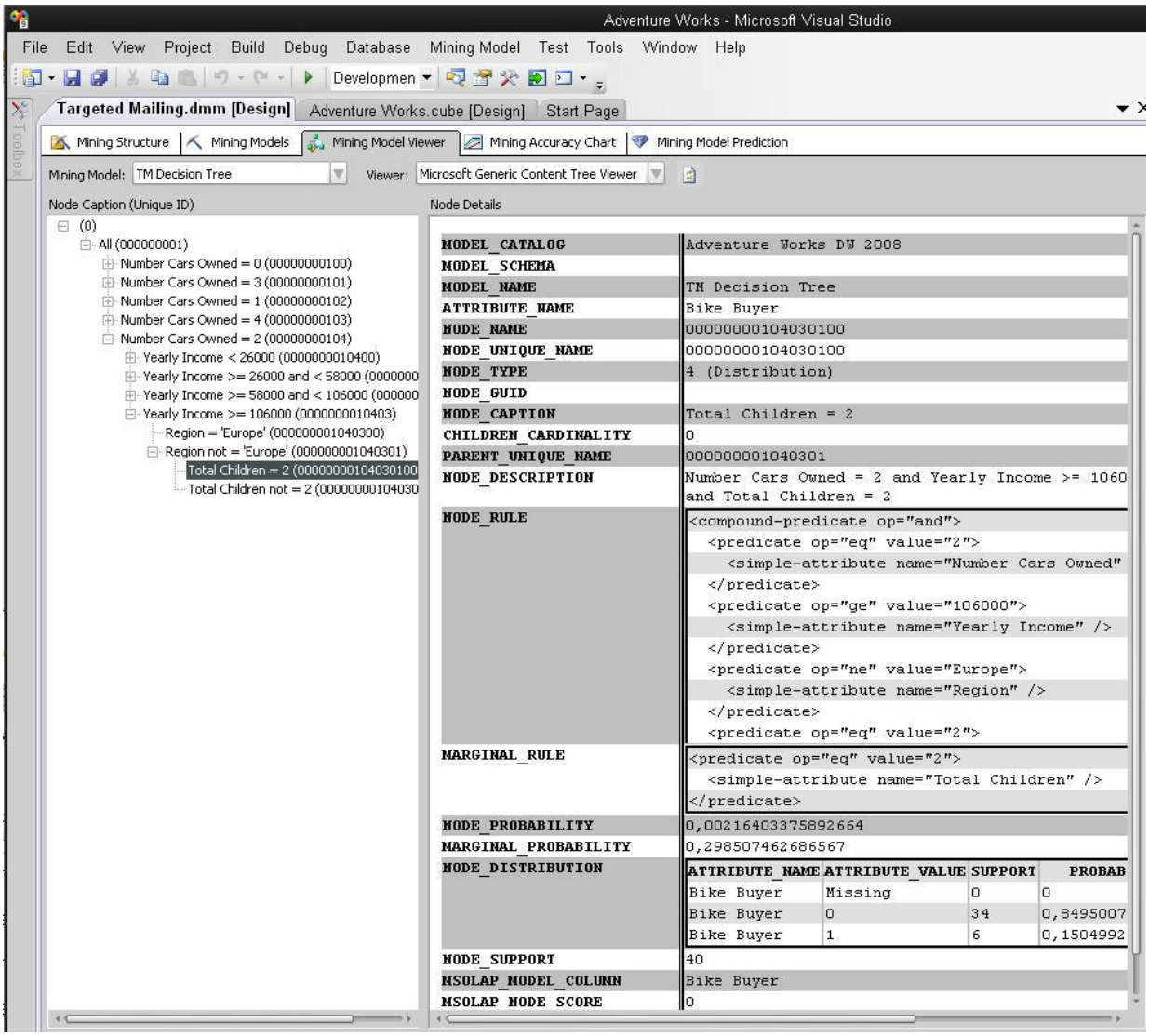
Sljedeći tip ugrađenog prikaza Microsoft Tree Viewer-a za ovaj algoritam je Dependency Network prikaz. Taj prikaz omogućuje brzo saznanje o tome koji podaci su u najjačoj vezi s određenim čvorom. Moguće je prilagoditi jačinu prikazane povezanosti pomicanje klizača s lijeve strane dijagrama, gore ili dolje. Dependency Network prikaz prikazan je na slici 3-14, na kojoj je vidljivo da su najjače povezani faktori s kupnjom bicikla broj automobila, godišnji prihod, i regija.



Slika 3-14. Dependency Network za Microsoft Decision Trees algoritam

Isto kao i Decision Tree prikaz za Microsoft Tree Viewer, tako i Depencency Network prikaz uključuje neke mogućnosti podešavanja. Već navedeni klizač je naravno najzanimljiviji, osim toga moguće je zumirati i pomicati dane informacije na radnoj površini. Na dnu prikaza nalazi se legenda, koja pomaže pri razumijevanju prikazanih rezultata.

Za razliku jednostavnog prikaza kao što je Decision Tree, Generic Content Tree Viewer prikazan na slici 3-15, daje veliku količinu detalja. Prikazuje obrađene rezultate u redovima i kolonama. Kod nekih modela rudarenja ovaj prikaz uključuje i ugniježđene tablice. Prikaz sadrži numeričke podatke o stopama vjerojatnosti i varijancama.



Slika 3-15. Microsoft Generic Content Tree View-er za Microsoft Decision Trees

Sljedeći model rudarenja koji je uključen u ovu strukturu rudarenja je model temeljen na Microsoft Clustering algoritmu. Nakon odabira tog modela rudarenja (TM Clustering) sa drop-down liste, na raspolaganju su novi pregledi koje je moguće odabrati. Microsoft Cluster Viewer uključuje četiri različita prikaza, a to su: Cluster Diagram, Cluster Profiles, Cluster Characteristics, i Cluster Discrimination.

3.6 Mining Accuracy Chart tab

Mining Accuracy Chart tab koristi se za provjeru valjanosti odabranog modela s obzirom na stvarne podatke, kako bi se saznala točnost modela i koliko će dobro predviđati buduće vrijednosti. Ovaj alat je prilično složen i sadrži četiri taba: Input Selection, Lift Chart, Classification Matrix, i Cross Validation.

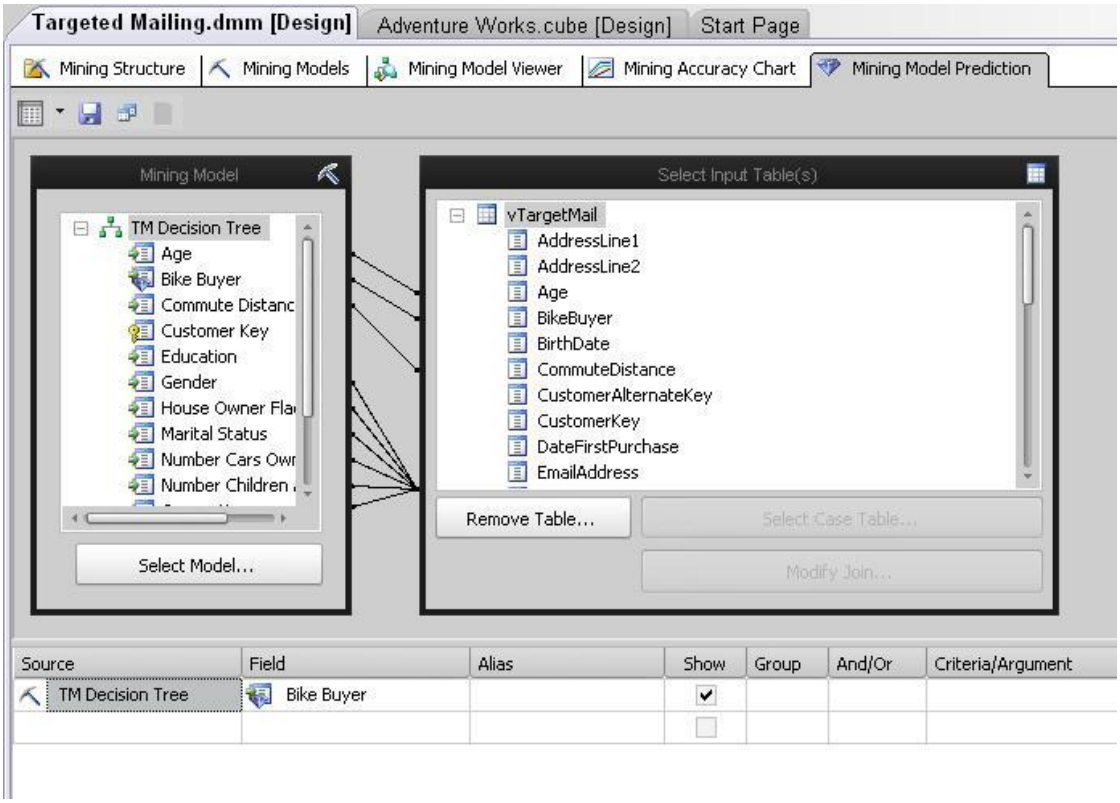
Rezultati Mining Accuracy Chart taba prikazuju se na više načina, uključujući Lift Chart, Classification Matrix, i Cross Validation. Uglavnom je moguće ocijeniti vrijednost rezultata koje određeni model predviđa. Ti rezultati mogu biti složeni za interpretaciju, pa će o ova tema biti posebno obrađena u poglavlju 3.9.

3.7 Mining Model Prediction tab

Ovdje je moguće napraviti predviđanja temeljena na odabranom modelu rudarenja sa novim eksternim podacima. Rad u ovom sučelju je zapravo pisanje (izrada) DMX upita za predviđanje korištenjem ponuđenih alata. DMX jezik sadrži nekoliko vrsta upita i ključnih riječi koje su uključene u ovu vrstu upita.

Kada se prvi put otvori sučelje, u prozoru Mining Model smješten je prvi model rudarenja podataka. Modele je moguće mijenjati klikom na tipku Select Model. Sljedeći korak je odabir novog izvora podataka. To se ostvaruje pritiskom na tipku Select Case Table. Nakon odabira tablice, BIDS automatski povezuje kolone izvora i destinacije sa istim nazivima. Za provjeru automatskog mapiranja potrebno je odabrati opciju Modify Connections nakon desnog klika na radnu površinu unutar Select Input Table prozora.

Nakon odabira tablice i provjere mapiranja prelazi se na izradu DMX upita. DMX upite je moguće pisati i pokretati pomoću SSMS-a. Slika 3-16 prikazuje BIDS sučelje Mining Model Prediction taba.



Slika 3-16. Mining Model Prediction tab omogućuje izradu DMX upita za predviđanje

Ovo je kraj poglavlja o sučelju BIDS-a za rudarenje podataka. Ovdje nisu prikazane sve njegove mogućnosti. U sljedećem poglavlju slijedi detaljan opis mogućnosti pojedinog algoritma. Razumijevanje tih algoritama je ključno za uspješnu implementaciju SSAS rudarenja podataka.

3.8 SQL Server 2008 algoritmi za rudarenje podataka

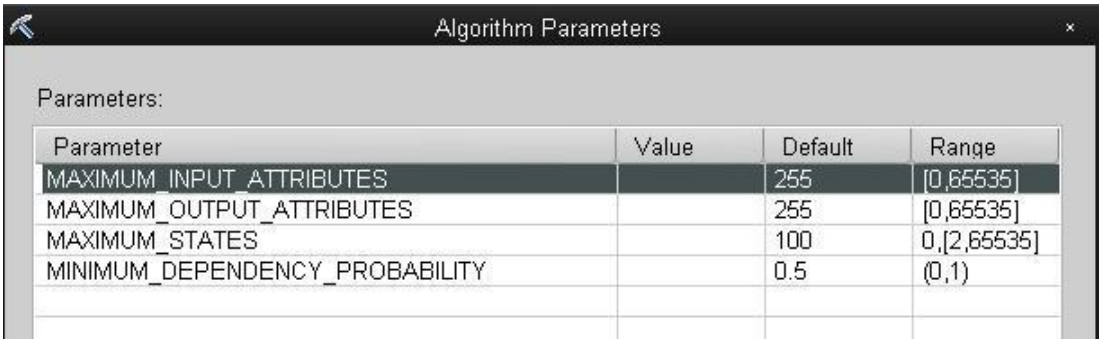
U ovom poglavlju obrađene su mogućnosti svih algoritama uključenih u ovaj softverski paket. Opisani su redom, od najjednostavnijeg do najsloženijeg. Za svaki algoritam navedene su njegove mogućnosti, konfiguracija, i neka napredna svojstva koja je moguće podešavati. Postoje dvije vrste algoritama, s nadzorom (engl. supervised) i bez nadzora (engl. unsupervised). Modeli rudarenja pod nadzorom zahtijevaju od korisnika odabir ulaznih kolona i kolona koje žele predvidjeti. Modeli rudarenja bez nadzora zahtijevaju od korisnika odabir samo ulaznih kolona. Ukoliko se model ne konfigurira na odgovarajući način, SSAS će prilikom izrade modela javiti grešku. Nenadzirani algoritmi su Clustering, Linear Regression, Logistic Regression, Sequence Clustering, i Time Series. Nadzirani algoritmi su Association, Decision Trees, Naïve Bayes, i Neural Network.

3.8.1 Microsoft Naïve Bayes

Microsoft Naïve Bayes je jedan od najjednostavnijih algoritama dostupnih u SSAS-u. Često se koristi kao polazna točka za razumijevanje osnovnih grupacija podataka. Ovaj tip obrade općenito je karakteriziran kao klasifikacija. Algoritam ima naziv naïve iz razloga što niti jedan od atributa nema veće značenje od drugog. Nazvan je po Thomasu Bayesu, čovjeku koji je osmislio način primjene načela matematike (vjerojatnosti) radi razumijevanja podataka. U ovom su algoritmu svi atributi tretirani nezavisno, odnosno nisu međusobno povezani. Algoritam radi tako da doslovno broji korelacije između atributa. Iako se može koristiti za predviđanje i grupiranje, najčešće se koristi tijekom ranih faza razvoja modela. Češće se koristi za grupiranje nego za predviđanje vrijednosti. Obično se svi atributi označuju kao ulazni ili za predviđanje, jer to od algoritma zahtijeva da ih sve uzme u obzir.

Naïve Bayes se može koristiti prilikom rada s velikom količinom podataka o kojima se zna malo. Primjerice, kompanija stekne podatke prilikom kupnje konkurentskog poduzeća. Za početak je najbolje koristiti Naïve Bayes algoritam.

Ovaj algoritam može procjenjivati samo diskretne (Discrete ili Discretized) sadržaje, što je značajno ograničenje. Ako se odabere struktura podataka koja sadrži kolone kojima vrsta sadržaja nije označena kao Discrete (primjerice Continious), model rudarenja temeljen na Naïve Bayes algoritmu će ignorirati te kolone. Ovaj algoritam sadrži samo mali broj svojstava koja se daju podešavati. Za pregled parametara korišten je Target Mailing primjer. Nakon što je primjer otvoren u BIDS-u, potrebno je kliknuti na Mining Models tab. Zatim odabrati opciju Set Algorithm nakon desnog klika na model koji koristi Naïve Bayes algoritam. Na slici 3-17 prikazan je dijalog Algorithm Parameters.

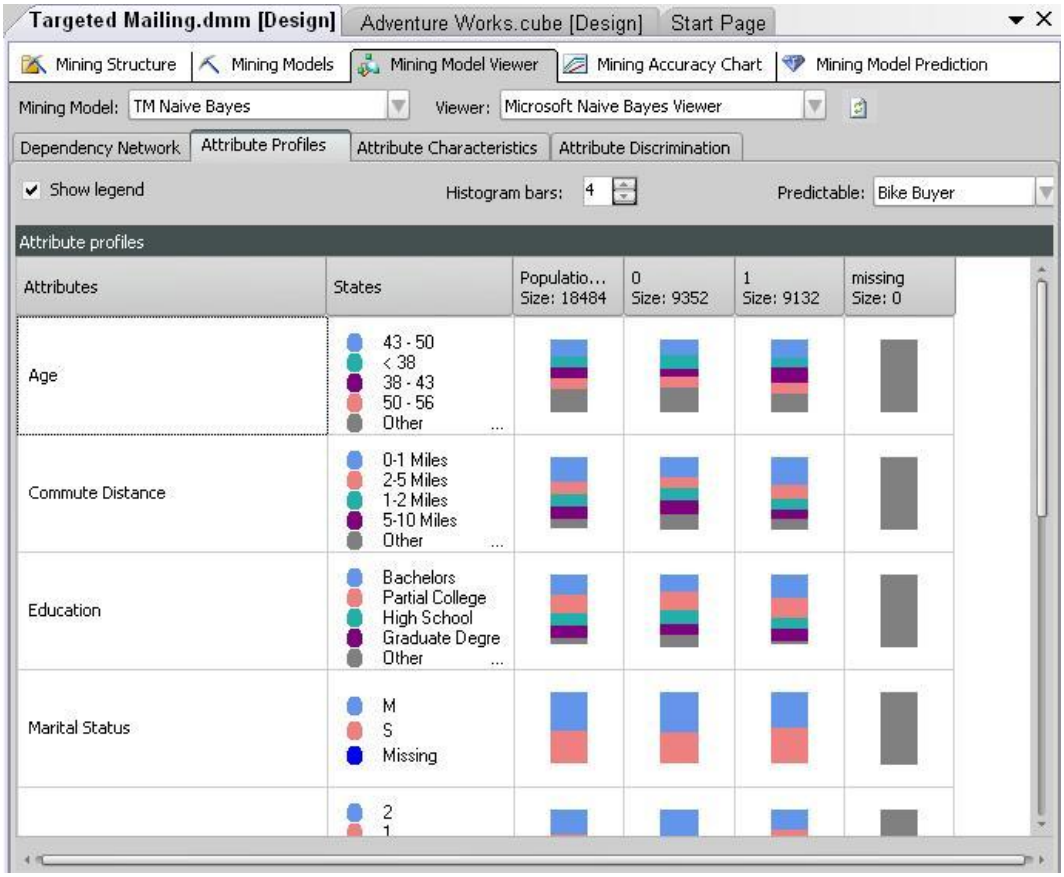


Slika 3-17. Dijalog Algorithm Parameters za Naive Bayes algoritam

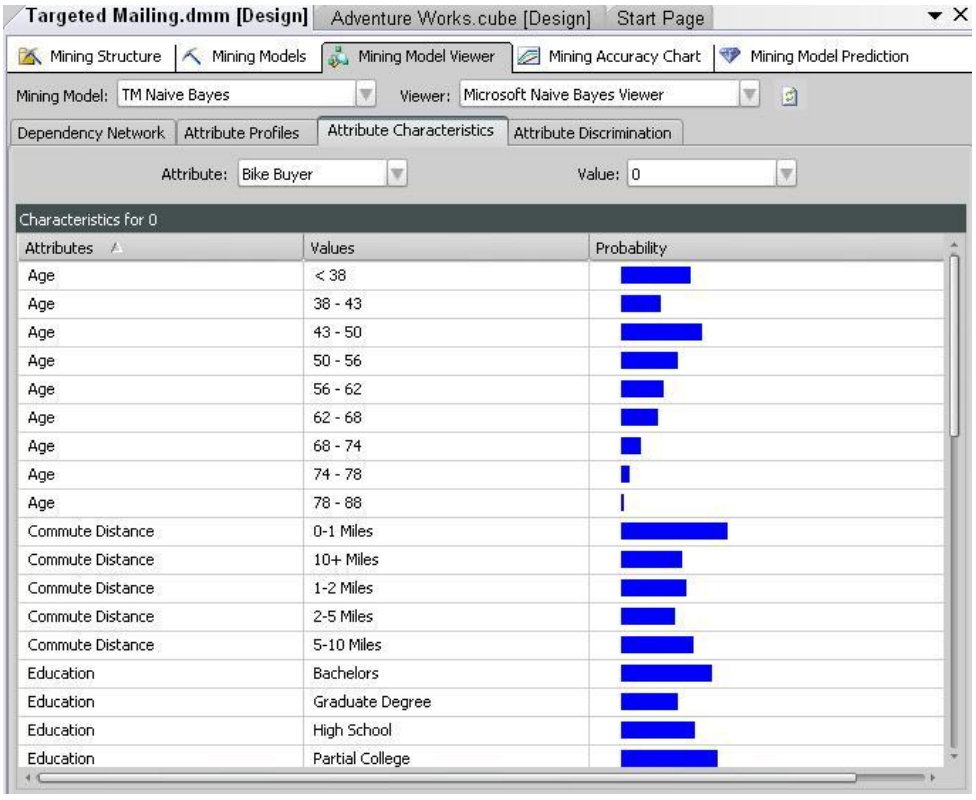
Kod Naïve Bayes algoritma moguće je konfigurirati sljedeće parametre: MAXIMUM_INPUT _ATTRIBUTES, MAXIMUM_OUTPUT_ATTRIBUTES, MAXIMUM_STATES, i MINIMUM _DEPENDENCY_PROBABILITY. Moguće je promijeniti zadane (engl. default) vrijednosti upisom novih vrijednosti u Value kolonu.

Iz razloga što se Naïve Bayes često koristi u projektima rudarenja podataka, osobito u ranoj fazi projekta, parametri se često podešavaju. Prva tri parametra su prilično očita. Podešavanje vrijednosti da bi se smanjio maksimalan broj ulaznih vrijednosti, izlaznih vrijednosti, i mogućeg grupiranja. Posljednji je manje očit. Kada se smanji ta vrijednost, smanjuje se s ciljem smanjenja broja čvorova ili grupa koje model stvara.

Microsoft Naïve Bayes Viewer koristi četiri vrste prikaza: Dependency Network, Attribute Profiles, Attribute Characteristics, i Attribute Discrimination. Dependency Network se često koristi jer ga je lako shvatiti pošto jednostavno prikazuje povezane atribute i snagu veze s odabranim čvorom. Taj pogled je prikazan na slici 3-14. Dio Attribute Profiles prikaza vidljiv je na slici 3-18, koja pokazuje kako je svaki ulazni atribut povezan sa svakim izlaznim atributom. Moguće je preuređivati redoslijed prikazanih atributa klikom na zaglavlje kolone u kojoj se nalazi željeni atribut i njegovim premještanjem. Moguće je mijenjati prikaz legende, promijeniti histogram, ili sakriti kolonu. Opcija ispod opcije Drillthrough nije dostupna, a to je iz dva razloga. Prvo, zato jer nije zadano da je Drillthrough uključen. Drugo, algoritam Naïve Bayes ne podržava Drillthrough.



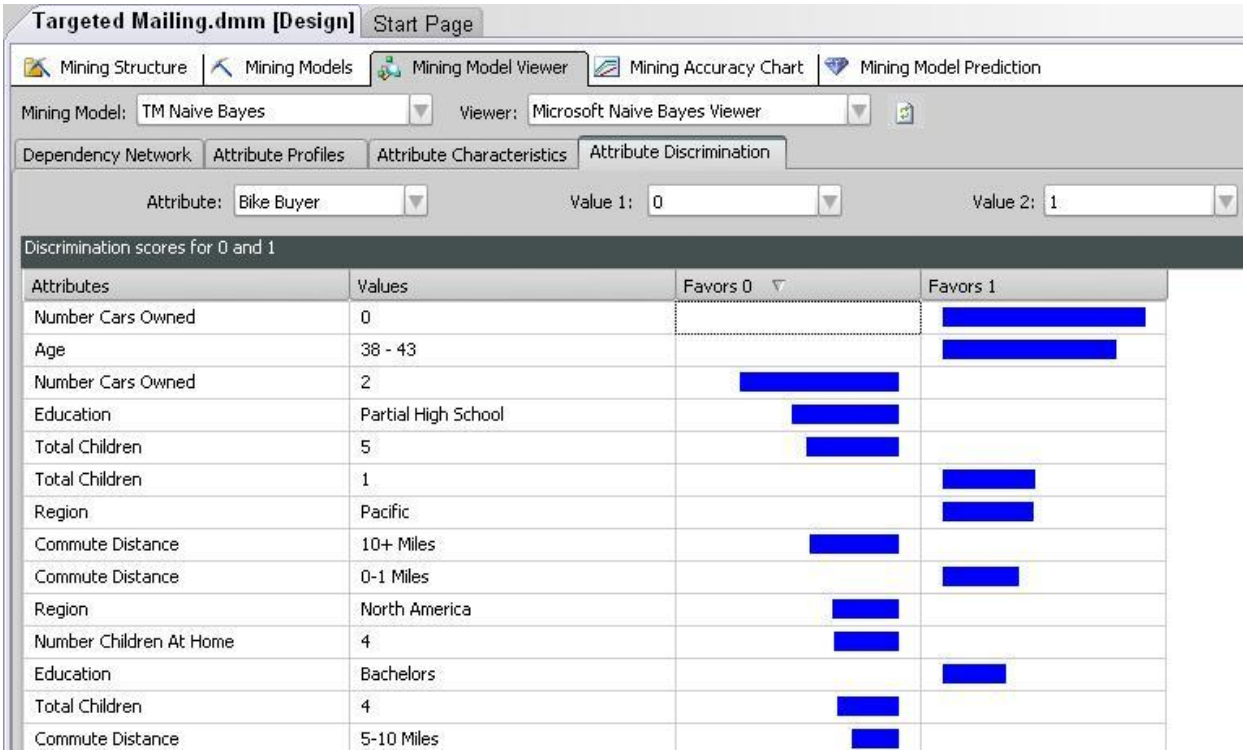
Slika 3-18. Attribute Profiles prikaz za Naive Bayes algoritam



Slika 3-19. Attribute Characteristics prikaz za Naive Bayes algoritam

Na sljedećem prikazu, Attribute Characteristics, moguće je vidjeti sve atribute povezane sa izlaznom vrijednosti. U ovom primjeru, zadana vrijednost stanja namještena je na 0, što znači da kupac ne kupuje bicikl. Zadana vrijednost sortiranja je od najjače do najslabije korelacije. Slika 3-19 prikazuje dio Attribute Characteristics prikaza. Sa slike je vidljivo da je kratka udaljenost do posla najviše povezana s ne kupnjom bicikla.

Pomoću Attribute Discrimination prikaza moguće je uspoređivati korelacije između atributa koji imaju dva različita stanja. Sa slike 3-20 je vidljivo da je vrijednost atributa posjedovanja 0 automobila u značajnijoj korelaciji s vrijednošću kupnje bicikla (Value 2: 1, što označuje kupnju bicikla), nego sa ne kupovinom bicikla (Value 1: 0, ne kupovina bicikla). Sljedeći značajni faktor je atribut Age sa vrijednošću 38-43. I ovdje, kao i u prethodnom prikazu, je moguće preslagivati rezultate klikom na zaglavlje kolona.



Slika 3-20. Attribute Discrimination prikaz za Naive Bayes algoritam

Kao što je već spomenuto, Naïve Bayes algoritam je jednostavan i često se koristi kao polazna točka za rudarenje podataka. Uključeni prikazi vrlo su jednostavni. Sljedeći algoritam je Microsoft Decision Trees.

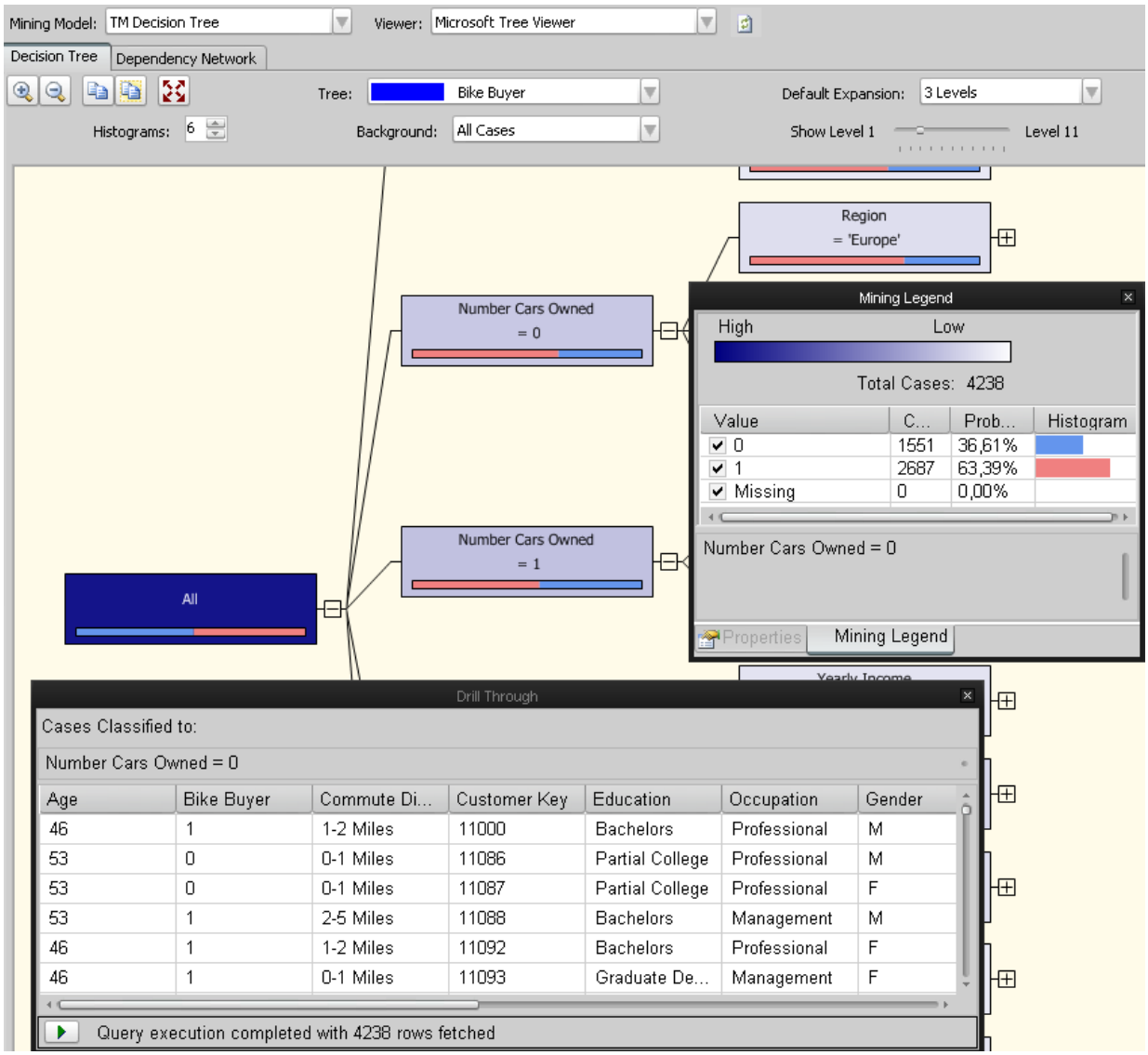
3.8.2 Microsoft Decision Trees

Microsoft Decision Trees je najčešće korišten algoritam. Djelomično zbog svoje fleksibilnosti, odnosno zato jer radi sa diskretnim i kontinuiranim atributima, te zbog toga što sadrži kvalitetne prikaze pomoću kojih je lako shvatiti rezultate. Algoritam se koristi za pregled i predviđanje. Također se koristi (obično zajedno s Microsoft Clustering algoritmom) za pronalaženje devijantnih vrijednosti. Microsoft Decision Trees procesuira ulazne podatke na način da ih dijeli na rekurzivne podskupove. Zadani preglednik rezultata ovog algoritma je struktura rekurzivnog stabla.

Ako se koriste diskretni podaci, algoritam identificira ulaze koji su najuže povezani s predvidljivim vrijednostima, odnosno ukazuje na kolone koje su od odabranih atributa najpredvidljivije. A ako se koriste kontinuirani podaci, algoritam koristi standardnu linearnu regresiju kako bi odredio gdje se stablo grana.

Slika 3-21 pokazuje Decision Tree prikaz. Svaki čvor sadrži oznaku koja ističe vrijednost. Klikom na čvor pojavljuju se detaljne informacije unutar Mining Legend prozora. Pomoću raznih drop-down lista je moguće konfigurirati prikaz, neke od njih su: Tree, Default Expansion, Background, itd. Moguće je i koristiti opciju Drilltrough, ako je ista prethodno uključena na modelu. Rezultat Drilltrough opcije za rezultat Number Cars Owned = 0 prikazan je na slici 3-21.

Microsoft Decision Tree je jedan od najčešće korištenih algoritama kod implementacije realnih projekata rudarenja podataka. Naročito se koristi za razradu marketinških scenarija kako bi se pronašli najuže povezani atributi. Koristi se i za provedbu zakona, odnosno za pronalaženje osobina ili atributa koji su najuže povezani s ponašanjem prijestupnika.



Slika 3-21. Decision Tree

Rezultate Microsoft Decision Tree algoritma moguće je poboljšati drugačijim grupiranjem podataka. To je moguće učiniti raznim ETL procesima ili pomoću standardnih upita na izvor podataka prije stvaranja strukture rudarenja podataka. Važno je paziti na razinu procesuiranosti modela. Na to je moguće utjecati promjenom vrijednosti COMPLEXITY_PENALTY parametra u dijalogu Algorithm Parameters. Podešavanjem tog broja mijenja se složenost modela, uglavnom se smanjuje broj razmatranih ulaza što rezultira manjom veličinom Decision Tree prikaza. Primjerice, vrijednost 0.5 daje od 1 do 9 atributa, dok vrijednost 0.9 daje od 10 do 99 atributa.

Sljedeća mogućnost ovog algoritma je prikaz rezultata u više stabala. Za to je potrebno zadati više od jedne kolone za predviđanje (ili ako ulazni podaci sadrže ugniježđenu tablicu za koju je zadano predviđanje), i onda algoritam kreira odvojena stabla za svaku kolonu za koju

je zadano predviđanje. Iz Tree drop-down liste, unutar Decision Tree prikaza, odabire se stablo se želi prikazati.

Dependency Network prikaz je također dostupan u Microsoft Decision Tree algoritmu. Izgleda i funkcionira na sličan način kao i sa Naïve Bayes algoritmom. I ovdje je pomoću klizača moguće uklanjati slabije povezane čvorove.

3.8.3 Microsoft Linear Regression

Microsoft Linear Regression je varijacija Microsoft Decision Tree algoritma, i radi kao klasična linearna regresija, odnosno povlači najbolji mogući pravac kroz niz točaka (kada su izvori najmanje dvije kolone kontinuiranih podataka). Ovaj algoritam računa sve moguće veze između vrijednosti atributa i daje potpunije rezultate nego druge metode (koje nisu rudarenje podataka) koje primjenjuju linearnu regresiju. Osim Key kolona, moguće je koristiti samo kolone koje sadrže kontinuirane numeričke podatke.

Ovaj algoritam se koristi za vizualizaciju veza između dva kontinuirana atributa. Realni primjer toga je pronalazak veza između fizičkih smještaja pojedine lokacije u dućanu i stope prodaje pojedine stavke. Rezultat algoritma je sličan rezultatima ostalih modela linearne regresije za pronalazak veza. Za razliku od većine metoda linearne regresije, Microsoft Linear Regression algoritam računa sve moguće veze između svih ulaznih setova podataka. Po tome je različit od ostalih modela linearne regresije, koji uglavnom koriste progresivne tehnike odvajanja izvora.

Parametri koji se mogu konfigurirati su maksimalni ulaz (ili izlaz) i FORCE_REGRESSOR. Ovaj algoritam se koristi za predviđanje kontinuiranih atributa. Prilikom korištenja ovog algoritma, jedan atribut se označuje kao regresor. Atribut koji je regresor mora biti označen i kao kontinuirani sadržaj. Taj atribut se koristi kao ključna vrijednost u regresijskoj formuli. Moguće je ručno definirati regresijsku kolonu pomoću FORCE_REGRESSOR parametra. Umjesto toga, moguće je postaviti DMX REGRESSOR zastavicu na odabranu kolonu.

3.8.4 Microsoft Time Series

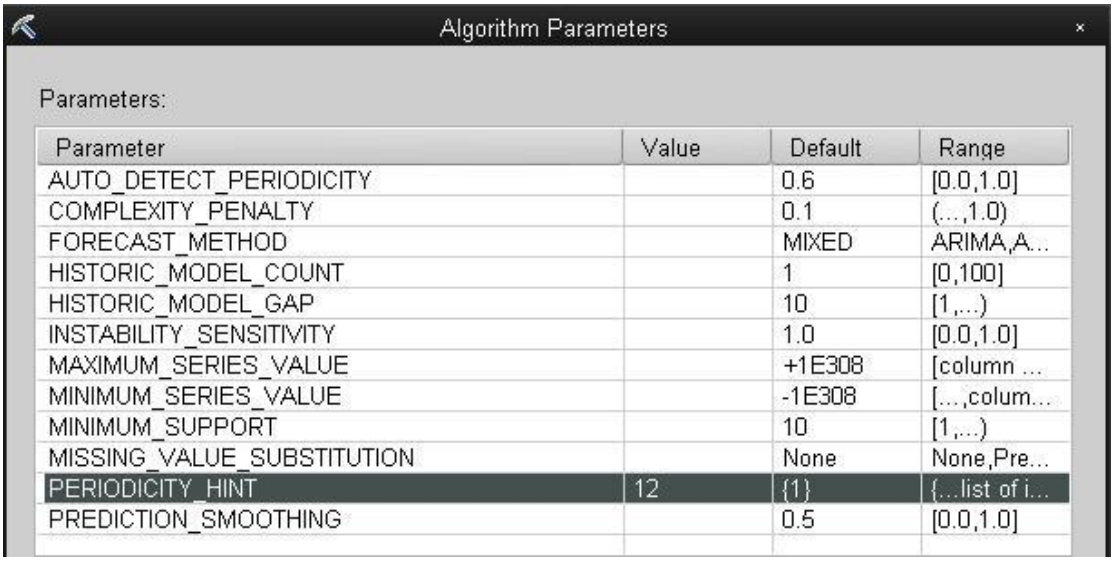
Microsoft Time Series algoritam se koristi za rješavanje učestalih poslovnih problema, odnosno za precizno prognoziranje. Često se koristi za predviđanje budućih vrijednosti, kao što je stopa prodaje određenog proizvoda. Ulazni atributi su najčešće kontinuirane vrijednosti.

Da bi se koristio ovaj algoritam, jedna kolona iz izvora podataka mora biti označena kao Key Time. Sve predvidljive kolone moraju biti kontinuiranog (Continious) tipa. Moguće je odabrati jednu ili više ulaznih kolona koje su predvidljive.

Moguće je konfigurirati stapanje dva algoritma vremenskih nizova. To se postiže konfiguracijom vrijednosti PREDICTION_SMOOTHING parametra. Kada su zadana dva algoritma oni se automatski stapaju. Moguće je i odabrati samo jedan od njih, a FORECAST_METHOD parametar pokazuje koji se algoritam koristi.

Kod Microsoft Time Series algoritma važno je uzeti u obzir odgovarajuću detekciju periodičkih uzoraka. Kada se razmatra periodičnost, potrebno je razumjeti sljedeće parametre:

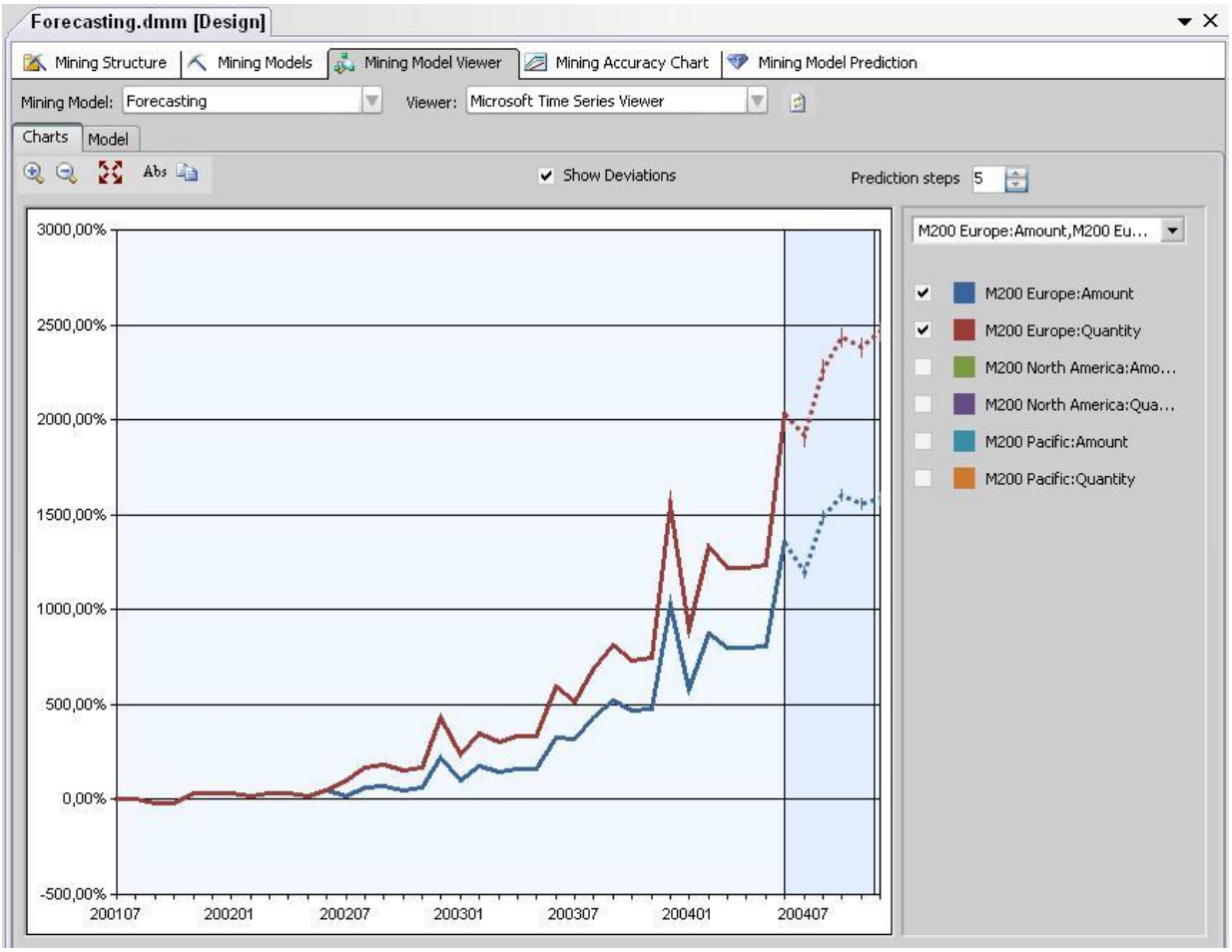
- AUTO_DETECT_PERIODICITY – Snižavanje zadane vrijednosti od 0.6 (u rasponu od 0 do 1.0) rezultira smanjenjem vremena procesuiranja modela jer se periodičnost otkriva samo na vrlo periodičnim podacima.
- PERIODICITY_HINT – Omogućen je unos više vrijednosti koje će uputiti algoritam u periodičnost podataka. Na slici 3-22 vidljivo je da je unesena vrijednost {12}. Primjerice, ako je periodičnost podataka godišnja ili kvartalna, trebalo bi unijeti {3, 12}



Slika 3-22. Konfiguracija parametara Microsoft Time Series algoritma

Microsoft Time Series Viewer pomaže u razumijevanju rezultata modela. Pokazuje predviđene vrijednosti s obzirom na odabrani vremenski niz. Moguće je zadati broj koraka predviđanja i prikazati devijacije, kao što je prikazano na slici 3-23. Prikazana je struktura

rudarenja podataka iz primjera Forecasting, i Forecasting model rudarenja u Charts prikazu, gdje su prikazani samo neki proizvodi. Prikaz prognoza ostalih proizvoda moguć je njihovim odabirom sa drop-down liste na desnoj strani prikaza. Ovaj model uključuje dvije predvidljive kolone, Amount i Quantity.

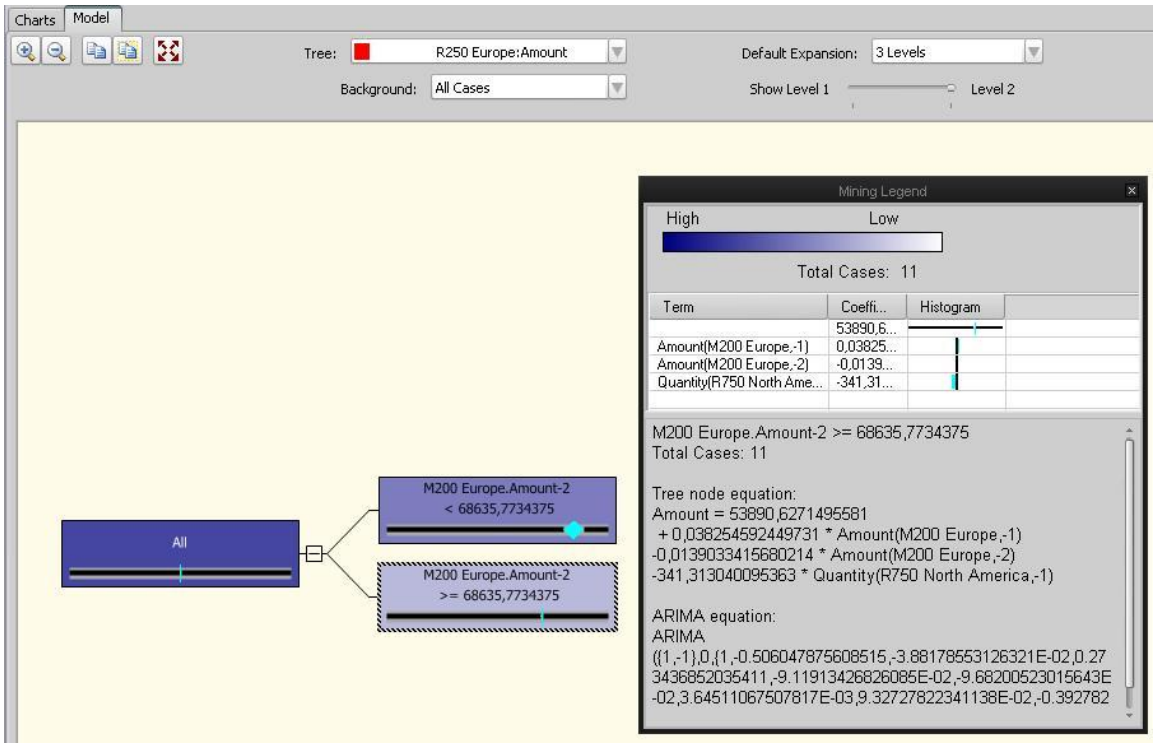


Slika 3-23. Charts prikaz - predviđene vrijednosti s obzirom na vremenski niz

U prikazanom pogledu odabrane su dvije izlazne vrijednosti Amount i Quantity za jedan proizvod (M200 Europe). Odabrana je i opcija Show Deviations kako bi se i te vrijednosti vidjele na grafu. Ako se miš zadrži na jednoj od linija na grafu, prikazuju se detaljne informacije o toj vrijednosti.

Sljedeći prikaz uključen u model rudarenja koji koristi Microsoft Time Series algoritam je Model prikaz. Izgleda slično kao Decision Tree prikaz koji se koristi kod modela rudarenja s Microsoft Decision Trees algoritmom. Međutim, iako modeli vremenskih nizova u Model prikazu imaju čvorove, kao i Decision Tree prikaz, Mining Legend prozor pokazuje informacije za ovaj algoritam, i to: koeficijente, histograme, i jednadžbe čvorova stabla.

Dostupan je pristup tim informacijama kako bi se bolje razumjela metoda pomoću koje su napravljena predviđanja. Model prikaz vidljiv je na slici 3-24.



Slika 3-24. Model prikaz pokazuje informacije o svakom čvoru

U prozoru Mining Legend moguće je vidjeti jednadžbe i koji je od dva dostupna vremenska algoritma (ARIMA ili ARTxp) korišten.

3.8.5 Microsoft Clustering

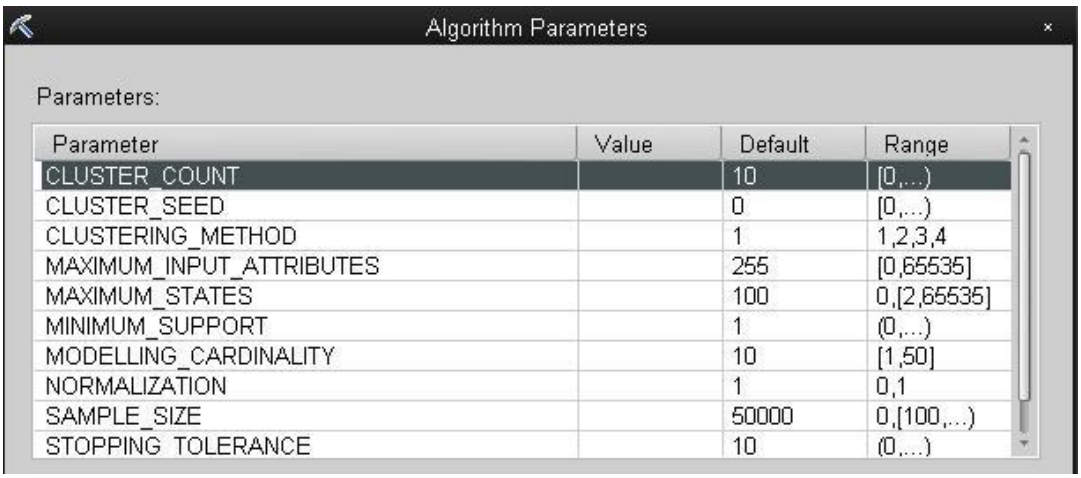
Microsoft Clustering algoritam služi za smislenu grupaciju izvornih podataka. Za razliku od Naïve Bayes algoritma koji zahtjeva ulaze diskretnog sadržaja i kojem su svi ulazni atributi jednake težine, Microsoft Clustering omogućuje veću fleksibilnost vrste ulaznog sadržaja i odabir metodologija grupiranja. Moguće je koristiti više vrsta sadržaja kao ulaze i odabrati metodu koja će se koristiti za izradu grupa.

Dakle , moguće je koristiti Continious, Discrete i većinu drugih vrsta sadržaja. Moguće je odabrati i predvidljivu vrijednost, označavajući kolonu kao Predict Only. Premda se Microsoft Clustering obično ne koristi za predviđanja.

Kod korištenja Microsoft Clustering algoritma bitno je razumjeti dostupne vrste grupiranja. Vrste grupiranja nazivaju se tvrde ili meke, a odabiru se pomoću CLUSTERING_METHOD parametara. Metode koje je moguće odabrati su: Scalable EM

(Expectation Maximization), Non-Scalable EM, Scalable K-Means, ili Non-Slalable K-Means. Scalable EM se smatra mekom metodom, a K-Means tvrdom zato što stvara grupe i zatim pridružuje podatke samo jednoj grupi bez da se preklapaju. EM grupiranje ima suprotan pristup, te su preklapanja dozvoljena. Skalabilnost modela odnosi se na to da li je u procesuiranje uzet cijeli set izvornih podataka, ili pak jedan njegov dio. Primjerice, zada se da je maksimalan broj redova u početnom procesu Scalable EM metode 50000 redova. Ako je taj broj redova dovoljan algoritmu da proizvede smislene rezultate, ostali redovi se ne uzimaju u obzir. Nasuprot tome, Non-Scalable EM metoda koristi cijeli set podataka u početnom procesu. Zato izvedba Scalable EM metode može biti i do tri puta brža nego od Non-Scalable EM. Skalabilnost kod K-Means metoda radi na istom principu. To znači da Scalable K-Means metoda naknadno, odnosno nakon prvog pokretanja, učitava dodatne redove ako joj prvih 50000 nije bilo dovoljno da proizvede smislene rezultate.

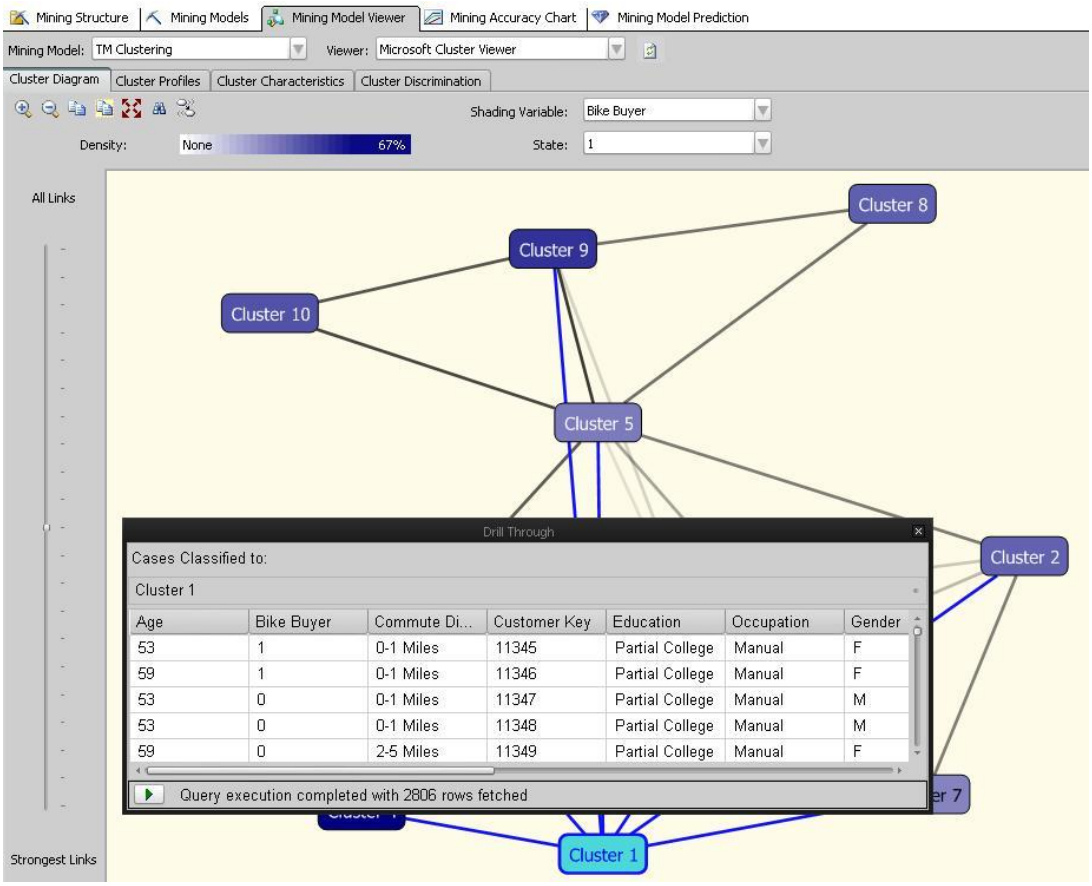
Vrijednosti na koje se može postaviti CLUSTERING_METHOD parametar kreću se između 1 i 4, a to znači da svaki broj predstavlja jednu od navedenih metoda. Na slici 3-25 vidljivo je da je zadana vrijednost 1. Vrijednosti za pojedinu metodu su: (1) Sclable EM, (2) Non-Scalable EM, (3) Scalable K-Means, (4) Non-Scalable K-Means.



Slika 3-25. Podešavanje parametara Microsoft Clustering algoritma

Nakon podešavanja modela, moguće je koristiti Microsoft Cluster View radi boljeg razumijevanja kreiranih grupa. Četiri su tipa Microsoft Cluster prikaza: Cluster Diagram, Custer Profiles, Cluster Characteristics i Cluster Discrimination. Kada je kod Cluster Diagram prikaza Shading Variable zadana za cijelu populaciju, nije moguće mijenjati vrijednost State opcije. Obično se Shading Variable vrijednost prilagodi nečemu drugome, kao što je to pokazano na slici 3-26 gdje je odabran Bike Buyer. Na slici je i State vrijednost postavljena

na 1. Sa slike su vidljivi i rezultati DrillThrough (Columns Only) opcije nad grupom Cluster 1. Na taj način moguće je koristiti rezultate Microsoft Clustering algoritma radi razumijevanja karakteristika kupaca bicikla.



Slika 3-26. Cluster Diagram prikaz daje informacije o varijablama i grupama

Više informacija o pojedinom klasteru moguće je dobiti zadržavanjem miša iznad istog. Moguće je i promijeniti naziv klastera desnim tipkom miša.

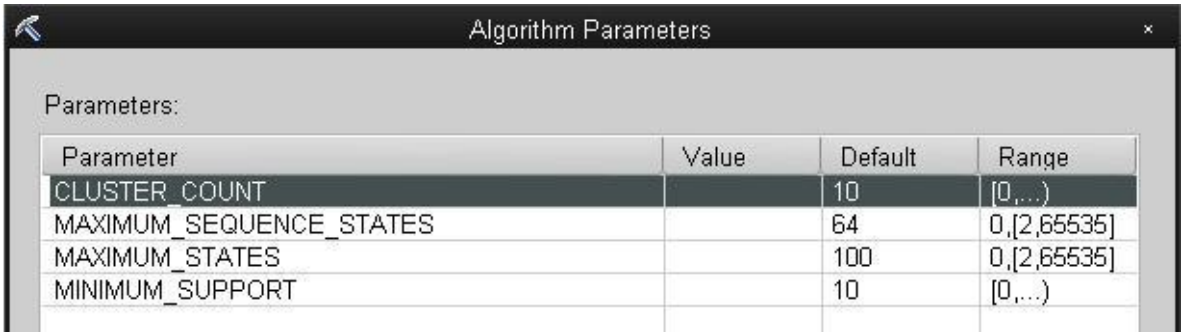
Preostala tri prikaza, vrlo su slična istoimenim prikazima dostupnim za pregled rezultata Naïve Bayes algoritma. Dakle, glavna razlika između Naïve Bayes i Microsoft Clustering algoritma je u tome što potonji pruža veću fleksibilnost vrsta ulaznih podataka i moguća je konfiguracija metoda grupiranja. Iz tih se razloga Microsoft Clustering koristi za slične zadatke, ali se češće koristi u kasnijim fazama projekata rudarenja podataka nego Naïve Bayes.

3.8.6 Microsoft Sequence Clustering

Microsoft Sequence Clustering algoritam daje slične rezultate kao i Microsoft Clustering, uz jedan bitan dodatak, a to je praćenje stanja između vrijednosti. Drugim riječima, on otkriva grupe određenog tipa, odnosno grupe sekvencijalnih podataka. Za provođenje ovog algoritma najmanje jedna kolona mora imati Key Sequence oznaku sadržaja, te ista mora biti dio ugniježđene tablice. Ako struktura izvora podataka ne uključuje odgovarajući tip podataka, ovaj algoritam nije dostupan u drop-down listi kada se kreira ili dodaje novi model rudarenja. Primjer uporabe ovog algoritma je click-stream analiza navigacije kroz web sadržaj (engl. web pages) neke web stranice (engl. web site). Click-stream analiza promatra kojem web sadržaju je pristupljeno i kojim redosljedom.

Ovaj algoritam koristi EM (Expectation Maximization) metodu grupiranja. Ali umjesto da samo prebrojavanjem pronalazi veze, on određuje i udaljenost između svih mogućih sekvenci izvora podataka. Zatim te informacije koristi za izradu sekvencijalno rangiranih grupa rezultata.

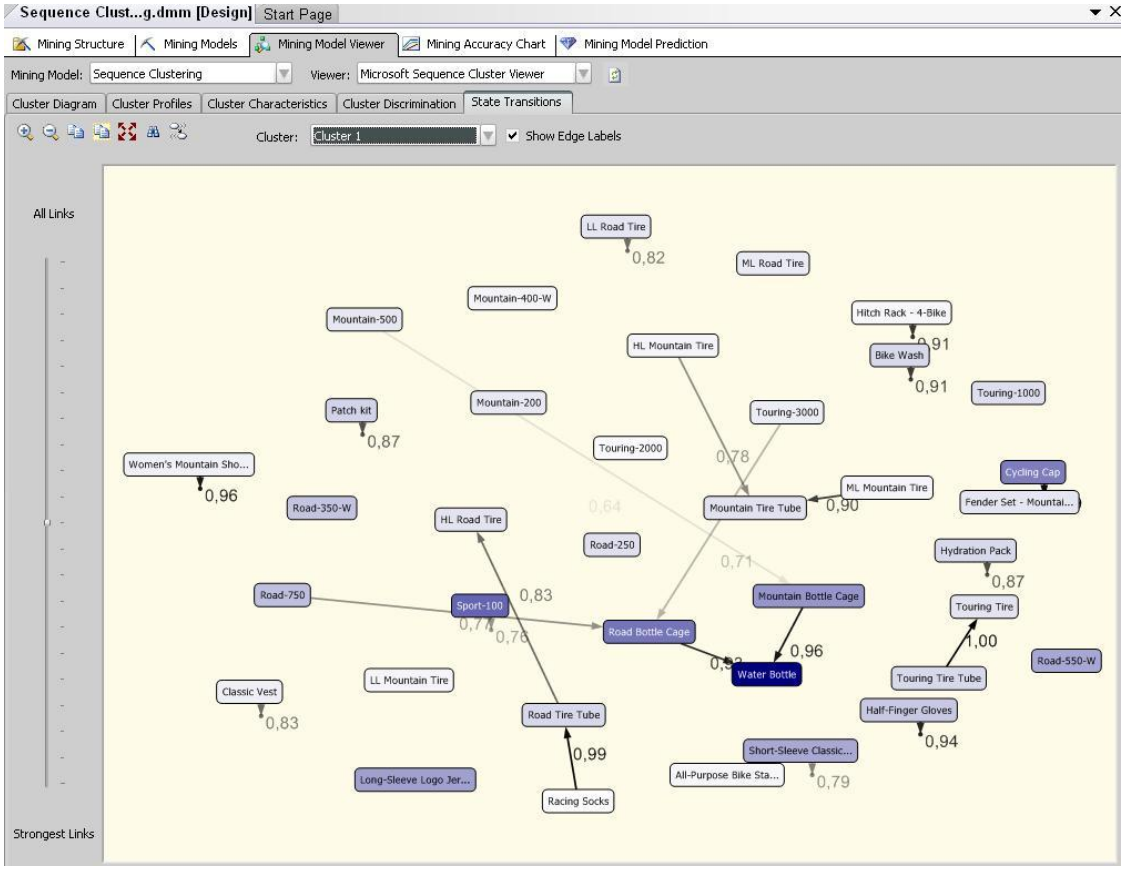
Zanimljiv parametar ovog algoritma je CLUSTER_COUNT, koji omogućuje podešavanje broja grupa koje algoritam izrađuje. Na slici 3-27 vidljivo je da je zadana vrijednost 10. Sljedeći parametar je MAXIMUM_SEQUENCE_STATES, zadana vrijednost ja 64, a moguće ju je promijeniti na vrijednosti između 2 i 65535. Promjenom tog parametra mijenja se maksimalni broj sekvencijalnih stanja koje će algoritam generirati.



Slika 3-27. Parametri Microsoft Sequence Clustering algoritma

Microsoft Sequence Cluster Viewer uključuje pet različitih načina prikaza Microsoft Sequence Clustering algoritma: Cluster Diagram, Custer Profiles, Cluster Characteristics, Cluster Discrimination i State Transition. Prva četiri prikaza vrlo su slični kao i kod Microsoft Cluster Viewer-a. Pomoću petog prikaza, State Transition prikaza, moguće je pogledati prijelazna stanja svih odabranih klastera. Svaki čvor (pravokutnik) prikazuje stanje modela.

Svaki čvor temelji se na vjerojatnosti prijelaza, koji su prikazani linijama između stanja. Ton boje predstavlja učestalost čvora u grupi. Kao i kod drugih prikaza grupa, i ovdje je standardno zadani prikaz cijelih populacija. Na slici 3-28 odabran je klaster Cluster 1. Broj prikazan pokraj čvora predstavlja vjerojatnost da utječe na povezani čvor.



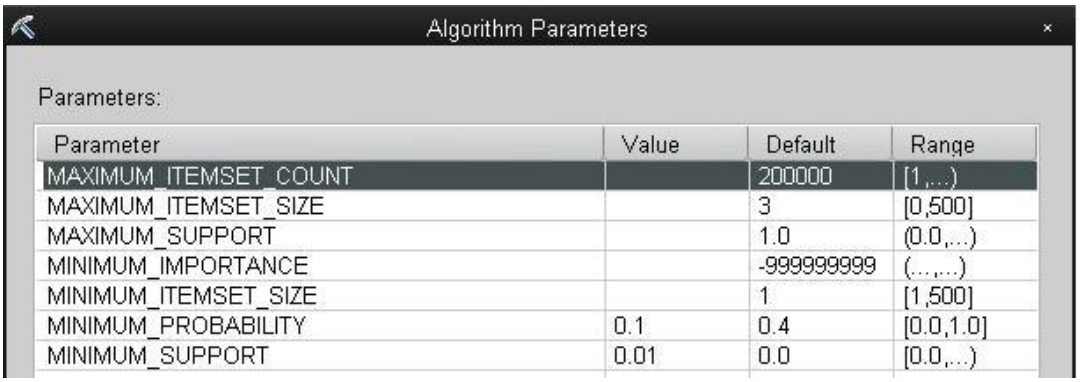
Slika 3-28. State Transition prikaz rezultata Microsoft Sequence Clustering algoritma

3.8.7 Microsoft Association

Microsoft Association algoritam se koristi za analiziranje grupa predmeta, zvanih skupovi predmeta, koji pokazuju zajedničke osobine. On stvara i dodjeljuje pravila tim skupovima predmeta. Ta pravila rangiraju vjerojatnost da će određeni skupovi predmeta biti zajedno. To se često naziva market-basket analiza. Izvor podataka može imati samo jednu predvidljivu vrijednost. Obično je to ključna (Key) kolona ugniježđene tablice. Sve ulazne kolone moraju biti diskretne (Discrete).

Ovaj je algoritam vrlo koristan trgovcima koji žele otkriti koje proizvode bi bilo dobro smjestiti zajedno, s ciljem povećanja prodaje. Naravno, to nije jedini slučaj u kojem se primjenjuje ovaj algoritam, ali je najčešći.

Procesuiranje izvora podataka, pronalaženje setova predmeta i pravila između njih je računalno vrlo zahtjevna. Kod promjena vrijednosti parametara ovog algoritma treba biti oprezan. Microsoft Association ima više podesivih parametara. Jedan od njih služi za podešavanje maksimalne veličine otkrivenih setova predmeta. Sa slike 3-29 je vidljivo da je zadana vrijednost tog parametra 3.

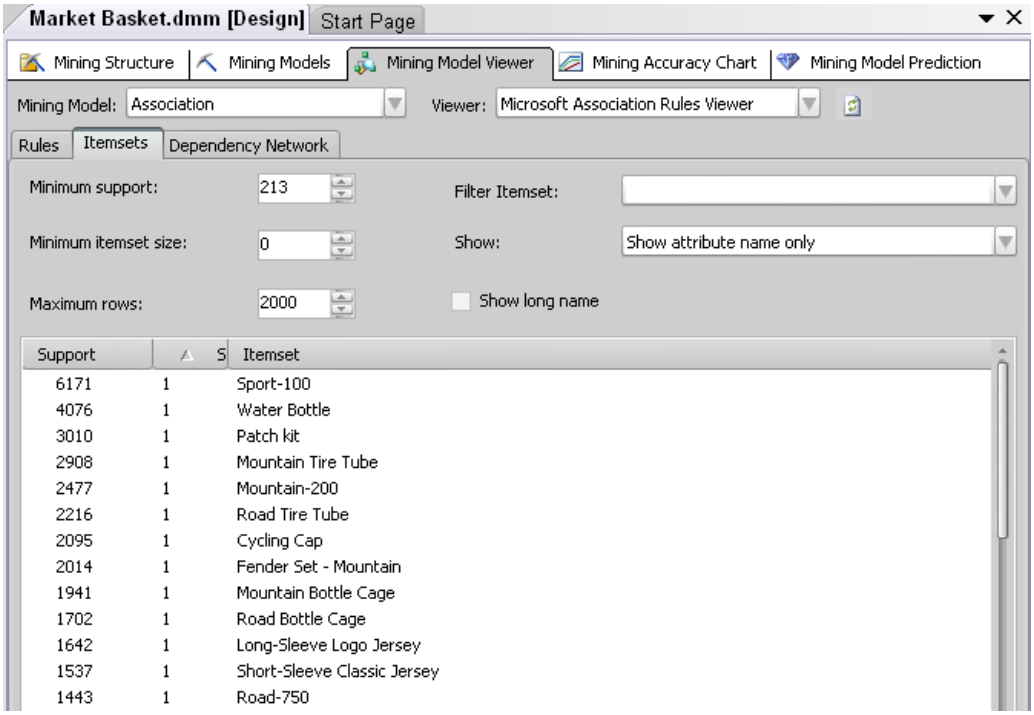


Slika 3-29. Parametri Microsoft Association algoritma

Microsoft Association Rules Viewer sadrži tri vrste prikaza za pregled rezultata Microsoft Association algoritma: Rules, Itemsets, i Dependency Network prikaz. Rules prikaz pokazuje listu pravila koja je algoritam otkrio. Obično je zadano da su pravila poredana po vjerojatnosti pojave. Naravno, moguće je promijeniti poredak klikom na bilo koje zaglavlje kolone unutar prikaza. Također je moguće prilagoditi prikaz da pokazuje pravila s različitim vrijednostima minimalnih vjerojatnost ili minimalnih važnosti konfiguracijom ponuđenih opcija.

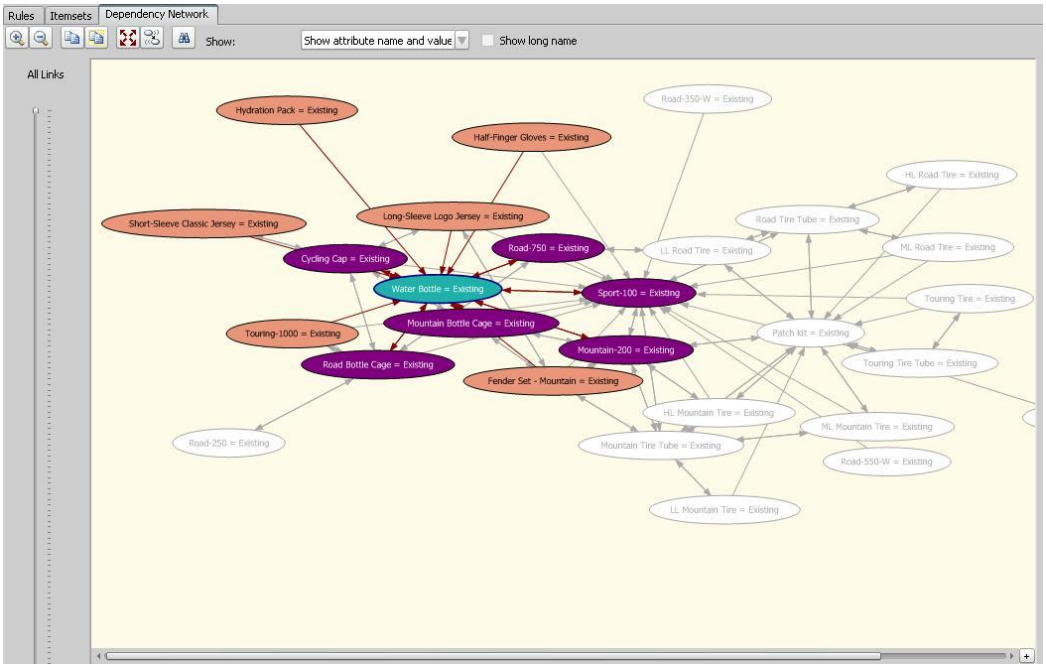
Itemsets prikaz omogućuje pregled otkrivenih setova predmeta. Ovaj prikaz je prikazan na slici 3-30, na kojoj su promijenjene neke od zadanih vrijednosti, kako bi se bolje razumjeli podaci. U Show drop-down listi odabrana je opcija Show Attribute Name Only. Zatim je kliknuto na kolonu Size da se podaci poredaju po broju predmeta u setu predmeta. Vidljivo je da je prvi od setova predmeta s jednim predmetom Sport-100 sa 6171 slučaja. I ovdje, kao i kod Rules prikaza moguće je podesiti Minimum Support. Također je moguće podesiti minimalnu veličinu seta predmeta kao i maksimalan broj prikazanih redova.

Moguće je i ručno napisati filtar za Rules i Itemset prikaz. Za to je potrebno upisati vrijednost u Filter Itemset rubriku. Primjerice, za ovaj primjer bi to moglo biti Mountain-200 = Existing. Nakon čega je potrebno pritisnuti tipku Enter kako bi se osvježio prikaz i prikazali filtrirani rezultati.



Slika 3-30. Itemsets prikaz pokazuje rezultate Microsoft Association algoritma

Za razumijevanje rezultata Microsoft Association algoritma najčešće se koristi Dependency Network prikaz. Taj prikaz pokazuje veze između predmeta i klizaćom smještenim s lijeve strane omogućeno je podešavanje prikaza. Klikom na određeni čvor, javljaju se boje koje ukazuju na to da li je čvor predviđen (od drugog čvora) ili je predvidio (neki čvor). Ovaj prikaz pokazan je na slici 3-31.

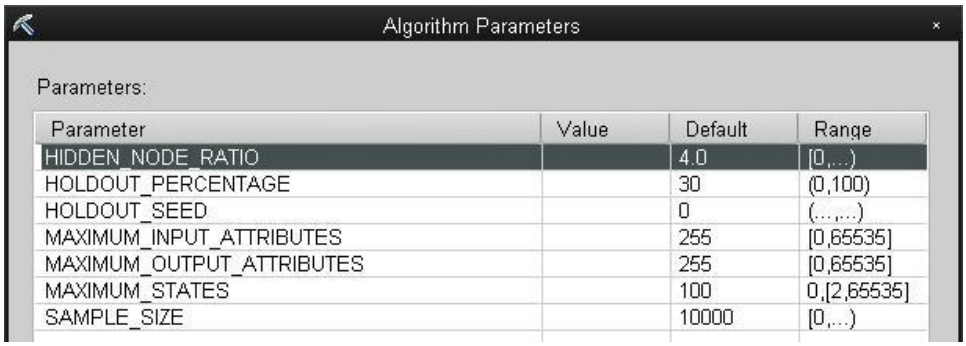


Slika 3-31. Dependency Network prikaz Microsoft Association algoritma

3.8.8 Microsoft Neural Network

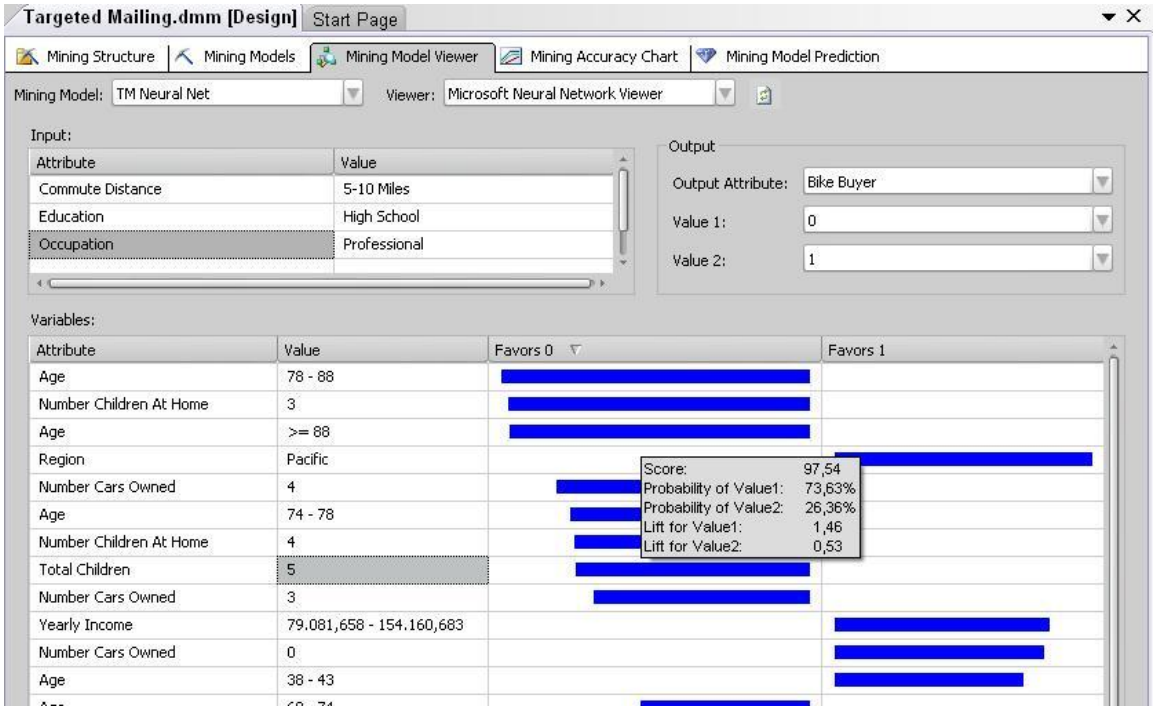
Microsoft Neural Network je daleko najsnažniji i najsloženiji algoritam. Ovaj algoritam kreira klasifikacijski i regresijski model rudarenja izgradnjom mreže neurona zvane Multilayer Perceptron. Slično kao i Microsoft Decision Trees algoritam, Microsoft Neural Network algoritam računa vjerojatnosti svakog mogućeg stanja ulaznog atributa kada je zadano svako stanje predvidljivog atributa. Kasnije se te vjerojatnosti mogu koristiti za predviđanje rezultata predviđenog atributa, s obzirom na ulazne atribute.

Ovaj algoritam je preporučeno koristiti kada ostali algoritmi ne uspijevaju dati smislene rezultate. Microsoft Neural Network može koristiti diskretne ili kontinuirane izvore podataka. Algoritam je potrebno dobro testirati na velikim izvorima podataka prije korištenja na produkcijskoj razini, to je zbog količine računalnih resursa koji su potrebni za provođenje ove vrste modela. Kod ovog, kao i kod drugih algoritama moguće je konfigurirati parametre pomoću Algorithm Parameters dijaloga. Parametre je preporučljivo mijenjati samo ako postoje poslovni razlozi, jer neke promjene rezultiraju zahtjevnijim i opsežnijim obradama podataka. Parametri su prikazani na slici 3-32.



Slika 3-32. Konfiguracija parametara Microsoft Neural Network algoritma

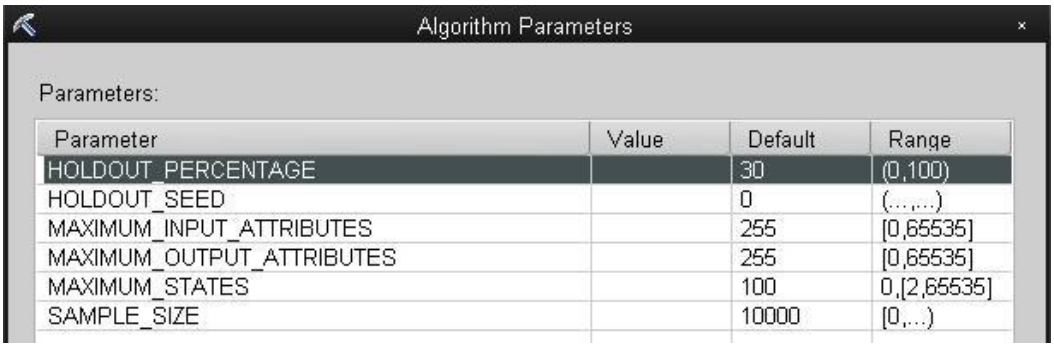
Microsoft Neural Network Viewer sadrži samo jedan prikaz za ovaj algoritam. Prikaz omogućuje dodavanje filtara ulaznih atributa, te podešavanje izlaznih atributa, kao što je prikazano na slici 3-33. Za ovu su sliku dodana tri filtara vezana uz Commute Distance, Education, i Occupation. Kada se miš zadrži iznad pojedine trake pojavljuju se oblačić s detaljnim informacijama.



Slika 3-33. Za Microsoft Neural Network algoritam postoji samo jedan prikaz

3.8.9 Microsoft Logistic Regression

Microsoft Logistic Regression je varijanta Microsoft Neural Network algoritma, odnosno algoritam koristi varijantu linearne regresije. Jedan primjer je dihotomija zavisne varijable, primjerice kao uspjeh/neuspjeh. Moguće je konfigurirati iste parametre kao i kod Microsoft Neural Network algoritma, osim HIDDEN_NODE_RATIO parametra koji je uvijek 0. Slika 3-34 pokazuje Algorith Parameters dijalog ovog algoritma.



Slika 3-34. Pregled parametara Microsoft Logistic Regression algoritma

Prikaz je isti kao i za Microsoft Neural Network algoritam, zato što je varijanta tog algoritma.

Kratak pregled svih algoritama

Slijedi kratki pregled devet algoritama za rudarenje podacima koji su uključeni u SSAS:

- **Microsoft Naïve Bayas algoritam** – Vrlo općenit algoritam, često se koristi kao polazna točka pri izradi projekata.
- **Microsoft Association algoritam** – Koristi se za provođenje market-basket analiza.
- **Microsoft Sequence Clustering algoritam** – Koristi se za provođenje sekvencijalnih analiza, kao što su click-stream analize (navigacija web stranica).
- **Microsoft Time Series algoritam** – Koristi se za prognoziranje budućih vrijednosti tijekom određenog vremenskog razdoblja.
- **Microsoft Neural Network algoritam** – Računalno vrlo zahtjevan, koristi se kao zadnja opcija, kad ostali algoritmi ne generiraju smislena rješenja.
- **Microsoft Logistic Regresion algoritam** – Koristi se kao alternativa tabličnoj logističkoj regresiji, fleksibilniji je.
- **Microsoft Decision Trees algoritam** – Najčešće korišteni algoritam.
- **Microsoft Linear Regression algoritam** – Varijacija Decision Trees algoritma.
- **Microsoft Clustering algoritam** –Za generalnu grupaciju, specifičniji od Naïve Bayes algoritma, koristi se za otkrivanje grupa povezanih atributa.

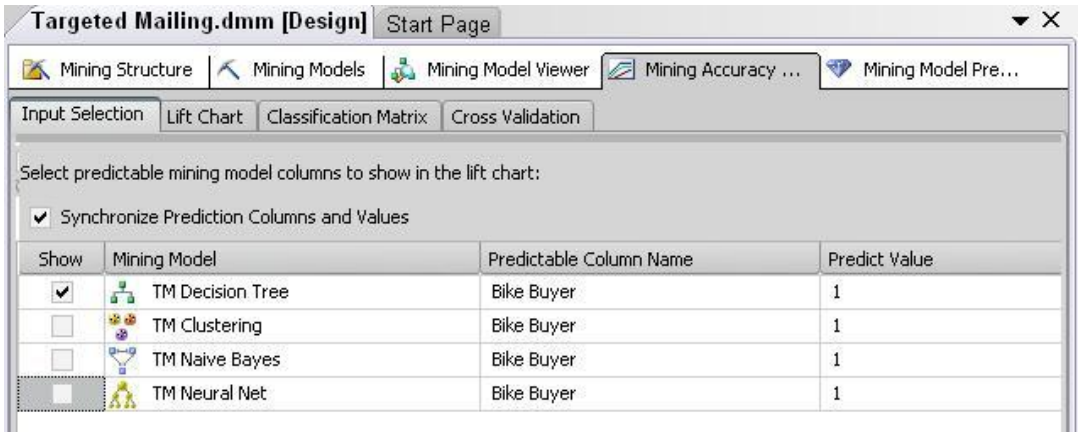
3.9 Validacija modela rudarenja

Neki korisnici ovog softvera ocjenjuju korisnost svojih model rudarenja podataka samo uporabom preglednika dostupnih u Mining Model Viewer tabu. Većina ipak provjerava i točnost tih modela, odnosno točnost njihovih rezultata. To se provodi radi daljnje prilagodbe određenog modela, ili da se odredi koji od algoritama najbolje odgovara određenoj poslovnoj situaciji.

Slijedi pregleda alata za validaciju koji su dostupni u BIDS-u. Svi oni nalaze se unutar Mining Accuracy Chart taba, a to su: Lift Chart, Classification Matrix, i Cross Validation.

Mining Accuracy Chart tab sadrži i pripremni Input Selection tab, čije je sučelje prikazano na slici 3-35.

Prije korištenja alata za validaciju potrebno je podesiti ulazne parametre. To se radi odabirom jednog ili više modela rudarenja iz odabrane strukture koju se želi analizirati, i odabirom predvidljive kolone i (opcionalno) predvidljive vrijednosti. Nakon toga, potrebno je odabrati izvor testnih podataka. Testni podaci sadrže točne odgovore (točne vrijednosti za predviđanja) i koriste se za ocjenu točnosti modela. Na dnu Input Selection taba moguće je promijeniti testni set podataka i odabrati bilo koji izvor koji je dostupan. Moguće je i definirati filtar za odabrani testni set podataka. Nakon konfiguracije ulaznih podataka klikom na Lift Chart tab kreira se grafikon za validaciju jednog ili više odabranih modela.



Slika 3-35. Input Selection tab

Kako bi se bolje shvatili rezultati Mining Accuracy Chart taba, potrebno je bolje razumijevanje Lift Lift Chart i Profit Chart grafikona.

3.9.1 Lift Chart

Lift Chart uspoređuje točnost svih ili samo odabranih predvidljivih vrijednosti modela sa prosječnom i savršenom predikcijom. Rezultat Lift Chart-a je grafikon koji sadrži liniju slučajnih pogodaka, odnosno 50 posto točnih vrijednosti, i liniju idealnih rezultata ili 100 posto točnih vrijednosti. Linija prosječnih vrijednosti prolazi sredinom grafa i dijeli ga na dva jednaka dijela, a linija idealnih vrijednosti pro prolazi gornjim dijelom grafa.

Lift Chart graf uspoređuje rezultate modela rudarenja sa podacima koji sadrže točne vrijednosti. Primjerice, ako algoritam vremenskih nizova predvidi rast određenih modela bicikla u nekom periodu, ti rezultati se uspoređuju sa stvarnom prodajom. Ti podaci sadrže točne predviđene vrijednosti za set podataka koji se promatra. Treba znati da graf pokazuje

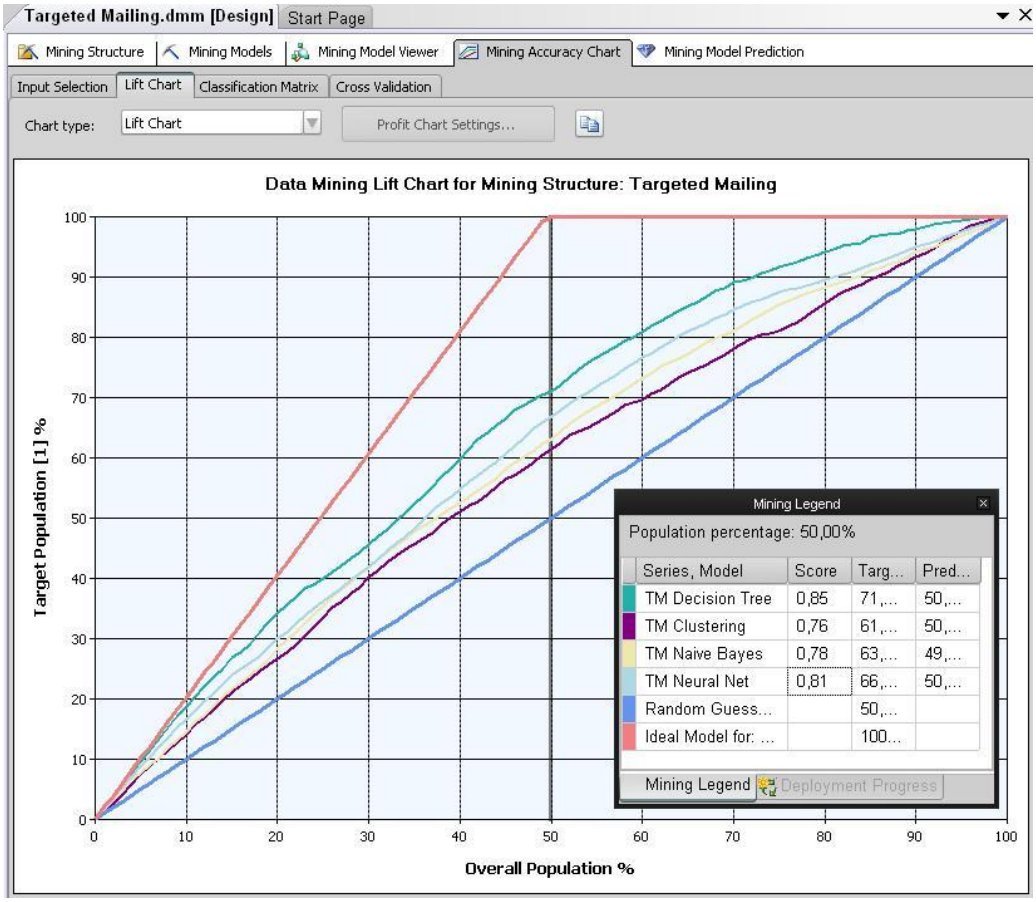
vjerojatnost predviđanja stanja i da nužno ne pokazuje točnost predviđanja stanja ili vrijednosti. Primjerice, može provjeriti da li model predviđa da li će netko kupiti bicikl ili neće. Lift Chart graf predviđa vrijednost samo onog ili onih stanja atributa koji su predodređeni za predviđanje (Predict).

Linija korisnih modela rudarenja podataka smještena je između linije prosječnih i idealnih vrijednosti. Što je linija modela bliže idealnoj liniji to je model efektivniji. Lift Chart koristi se za validaciju jednog ili više modela istovremeno. Ti modeli se mogu temeljiti na različitim algoritmima, istim algoritmom s različitim ulaznim parametrima, ili kombinaciji tih tehnika.

Na slici 3-36 prikazan je Lift Chart graf Targeted Mailing primjera strukture rudarenja podataka. Targeted Mailing struktura sastoji se od četiri modela rudarenja gdje je svaki model temeljen na različitom algoritmu rudarenja podataka. Svrha modela ove strukture je odrediti koji klijenti će najvjerojatnije kupiti bicikl. U ovom primjeru Lift Chart pokazuje koji model najbolje povezuje atribute s kupovinom bicikla. Takav prikaz pomaže pri odabiru modela na temelju čijih rezultata će se donositi odluke.

Izgled Lift Chart-a ovisi o vrsti sadržaja jednog ili više korištenih modela. Ako je model temeljen na kontinuiranim predvidljivim atributima, graf je više raspršen nego što je pravocrtan. Također, ako je model rudarenja temeljen na Time Series algoritmu, bolje je koristiti alternativne metode (napisati DMX upit) za validaciju nego Lift Chart.

Lift Chart sadrži i legendu rudarenja. Svaki redak u tablici pripada jednom modelu rudarenja, te po jedan za slučajni pogodak i idealni model. Legenda sadrži i tri kolone: Score, Target Population, i Predict Probability. Kolona Score pokazuje komparativnu vrijednost za uključene modele (veći brojevi su poželjni). Na slici 3-36 je vidljivo da model koji koristi Microsoft Decision Trees algoritam ima najveću vrijednost u Score koloni, i to je znak da najbolje od uspoređivanih modela predviđa ciljanu vrijednost, odnosno klijente koji će kupiti bicikl.

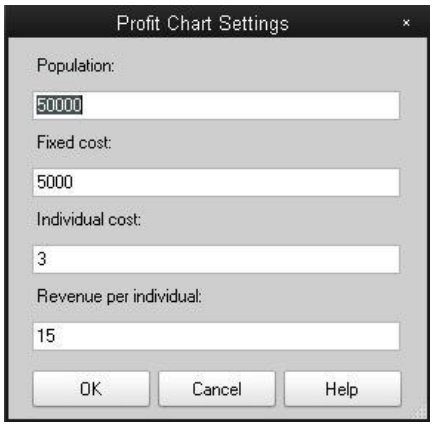


Slika 3-36. Lift Chart dozvoljava validaciju više modela rudarenja

Target Population kolona pokazuje koliko populacije će biti točno predviđeno za vrijednost zadanu na grafu. Vrijednost Overall Population je na grafu indicirana sivom vertikalnom linijom i postavljena je na 50 posto populacije, kao što je vidljivo na slici 3-36. Sa te slike je vidljivo i da za TM Decision Tree model ciljana populacija iznosi oko 71%, što je najveća vrijednost od analiziranih modela. To znači da je 71% klijenata, za koje algoritam predviđa da će prihvatiti ponudu, dobilo promidžbeni materijal. Predict Probability kolona pokazuje razinu vjerojatnosti potrebnu da svaka predikcija dostigne ciljanu populaciju. Dovoljno je reći da se sve tri vrijednosti legende koriste za donošenje najbolje poslovne odluke.

3.9.2 Profit Chart

Profit Chart pokazuje hipotetski porast profita s obzirom na svaki odabrani model. Kao što je spomenuto, Lift Chart i Profit Chart mogu se koristiti za sve algoritme osim za Microsoft Time Series algoritam. Moguća je konfiguracija Profit Chart-a klikom na Profit Chart Settings gumb. Ta akcija otvara dijalog prikazan na slici 3-37.



Slika 3-37. Dijalog Profit Chart Settings

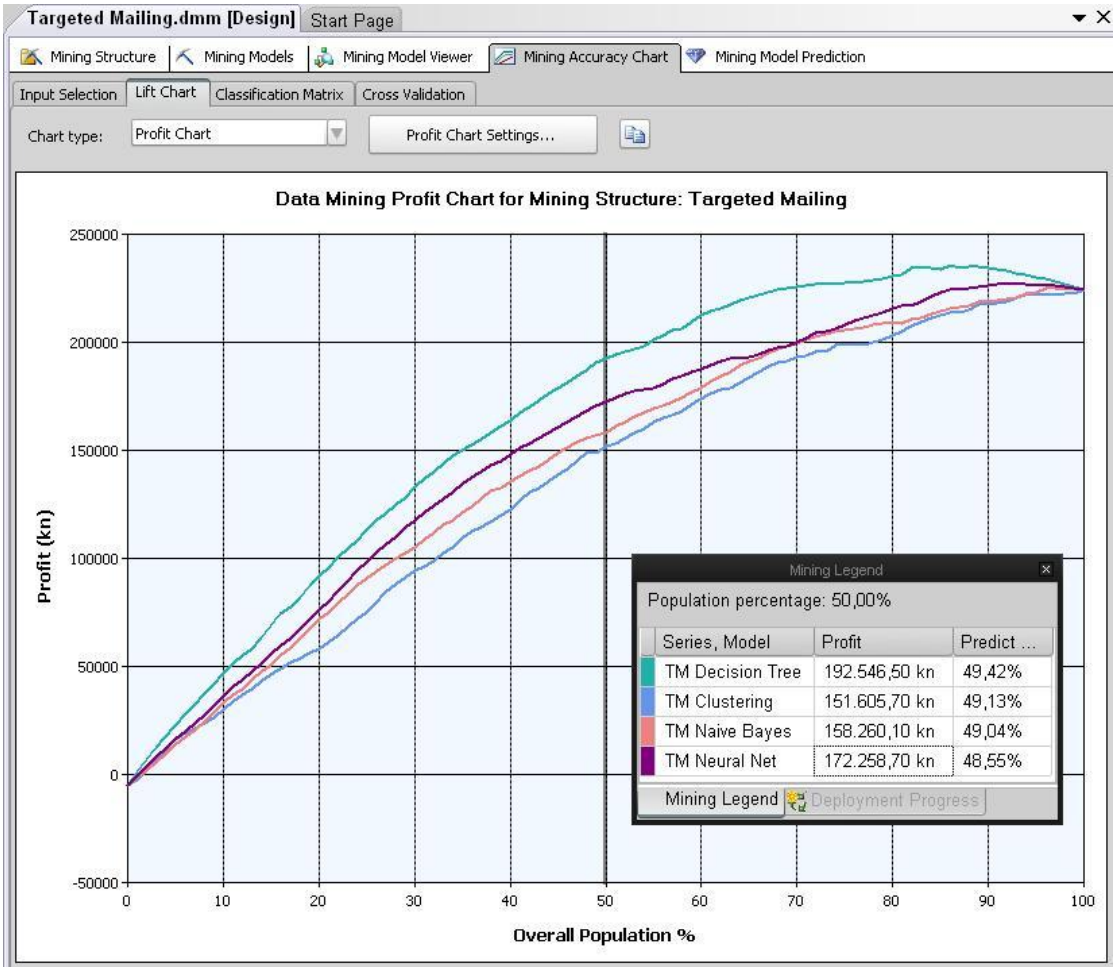
Konfigurirati se mogu sljedeće vrijednosti:

- Population – Ukupan broj slučajeva korištenih za izradu Profit Chart-a.
- Fixed Costs – Opći troškovi vezani uz određeni poslovni problem (primjerice, troškovi opreme i potrepština).
- Individual Cost – Troškovi pojedinog predmete (primjerice, troškovi tiskanja i slanja svakog kataloga).
- Revenue Per Individual – Procijenjeni prihod po svakom novom klijentu.

Nakon unosa tih vrijednosti, BIDS izrađuje graf koristeći DMX upite, baš kao i kod Life Chart-a.

Rezultati Profit Chart-a sadrže grafički prikaz i legendu. Graf prikazuje projiciranu dobit s obzirom na svaki uključeni model. Također je moguće koristiti legendu za pomoć pri evaluaciji učinkovitosti modela. Legenda sadrži red za svaki model. Za odabranu vrijednost na grafu, legenda pokazuje količinu projicirane dobiti. Pokazuje i predviđenu vjerojatnost za tu vrijednost profita. Ovdje su poželjniji veći brojevi.

Na slici 3-38 vidljivo je da se uporabom TM Decision Tree modela rudarenja podataka ostvaruje najveća dobit. Primjerice, na vrijednosti od 50 posto populacije projicirana vrijednost iznosi približno 192000 kn.



Slika 3-38. Profit Chart

Kao i kod Lift Chart-a, kod donošenja odluke trebalo bi uzeti u obzir i vrijednost predviđene vjerojatnosti. U ovom primjeru najveću vjerojatnost ima model TM Decision Tree (49,42%), dok najmanju ima TM Neural Network (48,55%). Te vrijednosti govore kolika je vjerojatnost da će se ostvariti prva vrijednost, odnosno profit.

3.9.3 Classification Matrix

Za velik broj validacije modela dovoljni su rezultati Lift i Profit Chart-ova. Međutim, u nekim situacijama potrebna je detaljnija validacija rezultata, odnosno kada su troškovi donošenja netočne odluke preveliki. Recimo da se radi o prodaji nekretnina, i potencijalni kupcima se nudi skupo putovanje na lokaciju nekretnine. Za taj slučaj potrebna je detaljna validacija pomoću klasifikacijske matrice.

Matrica je dizajnirana da funkcionira samo s diskretnim predvidljivim atributima. U tablici su prikazani rezultati koji pokazuju predviđene i stvarne vrijednosti za jedan ili više predvidljivih atributa. Ponekad su ti rezultati grupirani u sljedeće kategorije: netočno pozitivni

(false pozitive), točno pozitivni (true pozitive), netočno negativni (false negative), točno negativni (true negative). Tako da je mnogo specifičnija od Lift i Profit Chart-ova. Podnosi izvješća u točnim brojkama za sve četiri situacije. Kao i za Lift i Profit Chart-ove, set podataka za testiranje konfigurira se u Input Selection tabu.

Rezultat ove validacije je matrica ili tablica za svaki uključeni model rudarenja podataka. Svaka matrica pokazuje predviđene vrijednosti za model (u retku) i stvarne vrijednosti (u koloni). U tablici je vidljivo koliko je puta određeni model točno predvidio vrijednost. Classification Matrix analizira slučaj uključen u model s obzirom na predviđenu vrijednost, i pokazuje da li ta vrijednost odgovara stvarnoj vrijednosti.

Slika 3-39 prikazuje rezultate analizirane grupe modela. Slijedi komentiranje rezultata TM Decision Tree modela rudarenja. Ćelija u kojoj se nalazi broj 6317 označava broj točno negativnih za vrijednost 0. Zato što 0 označava broj kupaca koji nisu kupili bicikl, ova statistika govori da je model predvidio točnu vrijednost za kupce koji nisu kupili bicikl, u 6317 slučajeva. Ćelija koja sadrži broj 3035 označava broj netočno negativnih, odnosno koliko je puta model predvidio da bi netko kupio bicikl, kada zapravo nije. Ćelija koja sadrži broj 2225 označava broj netočno pozitivnih za vrijednost 1. Zbog toga što 1 znači da je kupac kupio bicikl, ova statistika govori da je u 2225 slučajeva model predvidio da netko ne bi kupio bicikl, kada ga je zapravo kupio. I na kraju ćelija koja sadrži broj 6907 označava broj točno pozitivnih za ciljanu vrijednost 1. Drugim riječima, u 6907 slučajeva model je točno predvidio da će netko kupiti bicikl.

Sumiranjem dijagonalno susjednih vrijednosti moguće je odrediti cjelokupnu točnost modela. Jedna dijagonala govori o ukupnom broju točnih predikcija, a druga o ukupnom broju pogrešnih predikcija. U ovom primjeru to bi izgledalo ovako. Za točne predikcije: 6317 + 6907 = 13224; i za netočne predikcije 3035 + 2225 = 5260. Proces je moguće ponoviti na svakom od uključenih modela, kako bi se odredio najprecizniji model.

Slika 3-39. Classification Matrix validacija

Dakle, za evaluaciju modela uključenih u Targeted Mailing strukturu moguće je napraviti tablicu koja izgleda kao tablica 3-2.

Tablica 3-2. Primjer rezultata Classification Matrix validacije

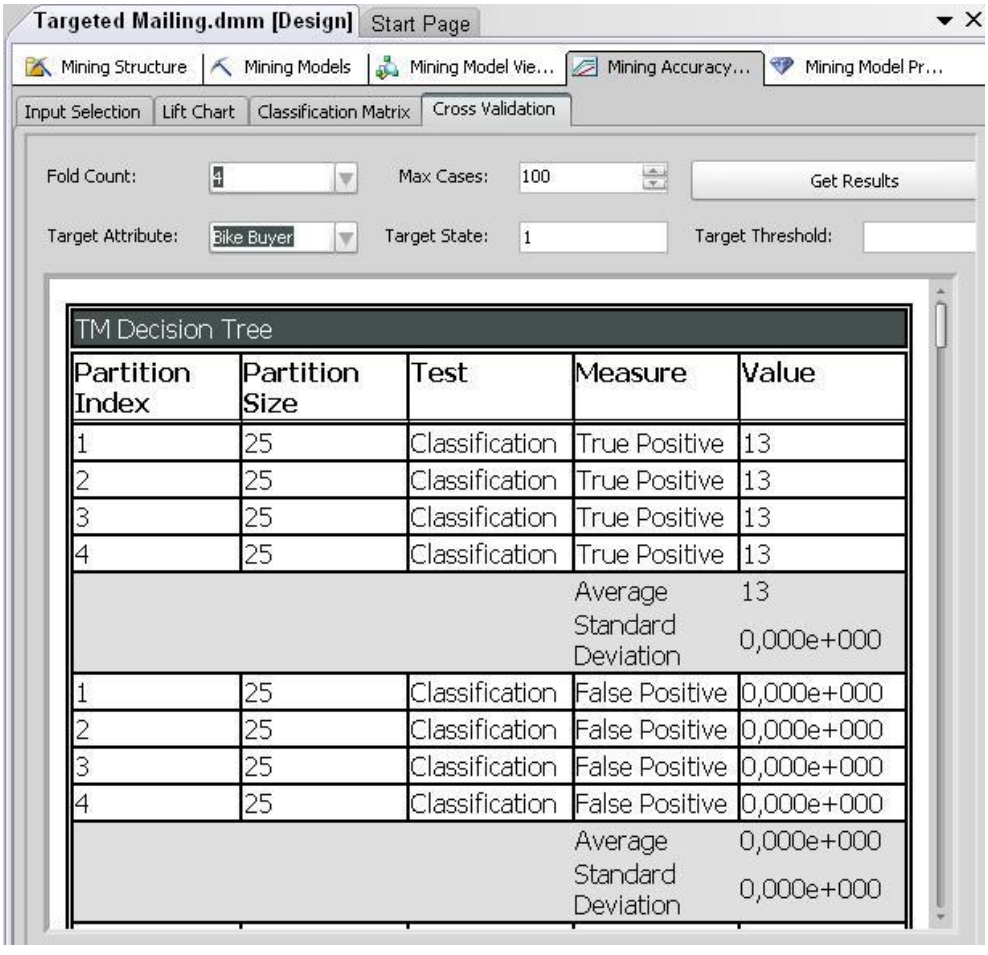
Model	Točni	Netočni
TM Decision Tree	13224	5260
TM Clustering	11275	7209
TM Naïve Bayes	11626	6858
TM Neural Network	12333	6151

Classification Matrix moguće je koristiti i za predvidljive vrijednosti koje imaju više od dva moguća stanja.

3.9.4 Cross Validation

Ovaj alat za validaciju radi malo drugačije od prethodno obrađenih. Cross Validation analiza ne zahtjeva odvojeni set podataka za treniranje i testiranje. Moguće je koristiti podatke za testiranje, ali to nije uvijek potrebno. Ova eliminacija potrebe za testnim (holdout) podacima čini Cross Validation prikladnijom za validaciju modela rudarenja podataka.

Cross Validation automatski dijeli izvor podataka na dvije jednake cjeline. Zatim izvodi iterativni test nad svakom cjelinom i prikazuje rezultate. Primjer prikaza rezultata pokazan je na slici 3-40. Cross Validation radi ovisno o vrijednosti definiranoj u Fold Count parametru. Standardna zadana vrijednost tog parametra je 10, što je jednako 10 setova. Maksimalna vrijednost tog parametra je 256. Naravno, veća vrijednost rezultira opsežnijim procesuiranjem. U korištenom primjeru korištene su sljedeće vrijednosti: Fold Count (4), Max Cases (100), Target Attribute (Bike Buyer), i Target State (1). Cross Validation je računski vrlo zahtjevna, i često je potrebno prilagoditi ulazne parametre za postizanje balansa između performansa i rezultata.



TM Decision Tree

Partition Index	Partition Size	Test	Measure	Value
1	25	Classification	True Positive	13
2	25	Classification	True Positive	13
3	25	Classification	True Positive	13
4	25	Classification	True Positive	13
			Average	13
			Standard Deviation	0,000e+000
1	25	Classification	False Positive	0,000e+000
2	25	Classification	False Positive	0,000e+000
3	25	Classification	False Positive	0,000e+000
4	25	Classification	False Positive	0,000e+000
			Average	0,000e+000
			Standard Deviation	0,000e+000

Slika 3-40. Cross Validation

Rezultati su prikazani na sličan način kao i kod klasifikacijske matrice, a to su grupe: točni pozitivni, netočni pozitivni, i tako dalje. Cross Validation se ne može koristiti za validaciju modela koji koriste Time Series ili Sequence Clustering algoritme.

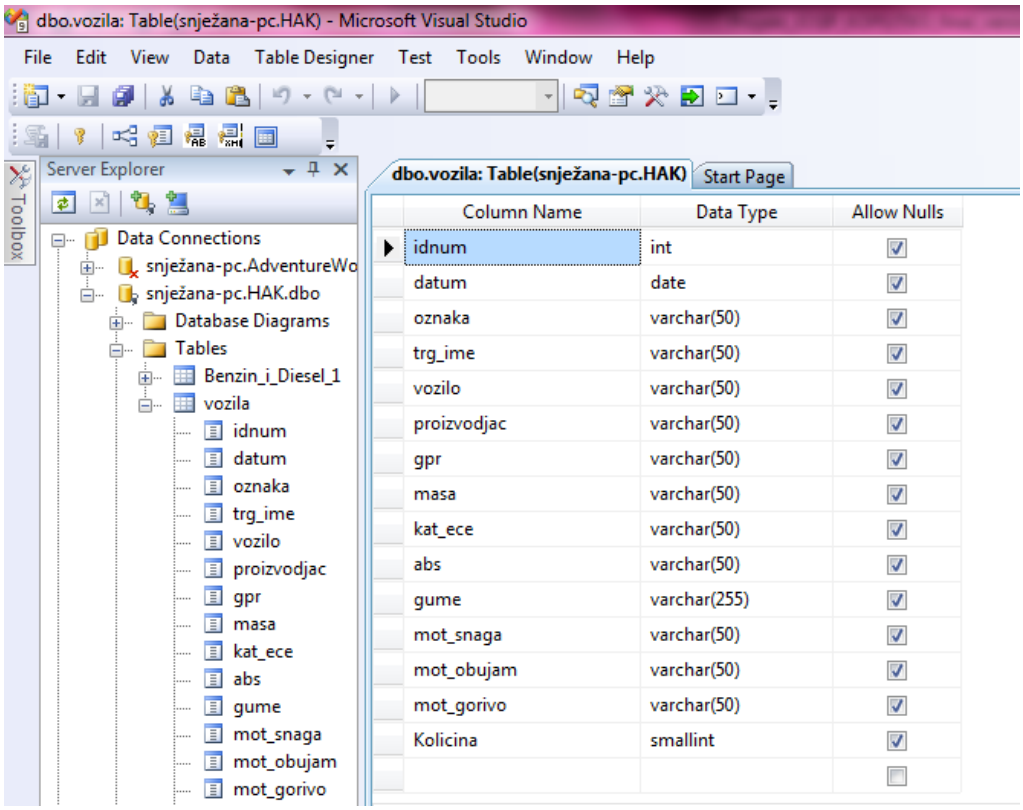
Nakon validacije modela, moguće se vratiti i napraviti neke promjene radi unaprjeđenja validacije, te ponovo provesti validaciju. Te promjene mogu uključivati izradu novih modela (koristeći druge algoritme), promjenu vrijednosti parametara algoritama, dodavanje ili uklanjanje izvornih kolona, ponovnom konfiguracijom vrijednosti atributa, i tako dalje.

4. Primjer primjene rudarenja podataka

U ovom poglavlju, kao primjer stvarne baze podataka sa stvarnim podacima, koristit će se HAK-ova baza podataka sa unaprijed definiranom strukturom i stvarnim podacima.

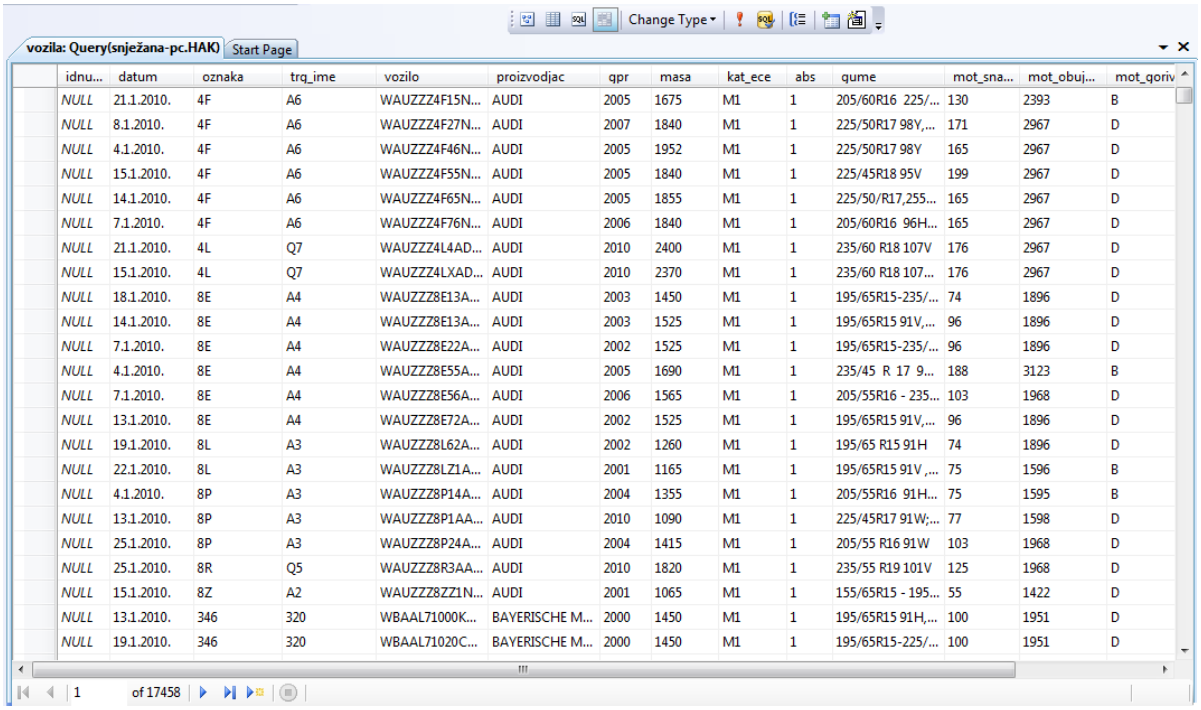
4.1 Definiranje strukture baze podataka

Baza podataka koja se koristi u ovom primjeru sadrži podatke o uvozu rabljenih vozila u Republiku Hrvatsku 2010. godine. Sastavljena je od jednog jedinog entiteta i to tablice po nazivom *vozila* u kojoj se nalaze sljedeći atributi: idnum, datum, oznaka,trg_ime, vozilo, proizvodjac, gpr, masa, kat_ece, abs, gume, mot_snaga, mot_obujam, mot_gorivo. Cijelokupna struktura baze podataka prikazana je na slici 4-1.



Slika 4-1. Definicija strukture HAK baze podataka

Na sljedećoj slici 4-2 prikazani su stvarni podaci tj. zapisi vezani uz odgovarajuće atribute kojima se oni ujedno i opisuju. Za ovakvu strukturu podataka može se reći da posjeduje i podatke i informacije, točnije, da su ti podaci prošireni u informaciju. Sljedeći logični korak jest da se te informacije transformiraju u znanje no više o tome u sljedećim poglavljima.



idnu...	datum	oznaka	trq_ime	vozilo	proizvodjac	qpr	masa	kat_ece	abs	qume	mot_sna...	mot_obuj...	mot_qoriv
NULL	21.1.2010.	4F	A6	WAUZZZ4F15N...	AUDI	2005	1675	M1	1	205/60R16 225/...	130	2393	B
NULL	8.1.2010.	4F	A6	WAUZZZ4F27N...	AUDI	2007	1840	M1	1	225/50R17 98Y,...	171	2967	D
NULL	4.1.2010.	4F	A6	WAUZZZ4F46N...	AUDI	2005	1952	M1	1	225/50R17 98Y	165	2967	D
NULL	15.1.2010.	4F	A6	WAUZZZ4F55N...	AUDI	2005	1840	M1	1	225/45R18 95V	199	2967	D
NULL	14.1.2010.	4F	A6	WAUZZZ4F65N...	AUDI	2005	1855	M1	1	225/50/R17,255...	165	2967	D
NULL	7.1.2010.	4F	A6	WAUZZZ4F76N...	AUDI	2006	1840	M1	1	205/60R16 96H...	165	2967	D
NULL	21.1.2010.	4L	Q7	WAUZZZ4L4AD...	AUDI	2010	2400	M1	1	235/60 R18 107V	176	2967	D
NULL	15.1.2010.	4L	Q7	WAUZZZ4LXAD...	AUDI	2010	2370	M1	1	235/60 R18 107...	176	2967	D
NULL	18.1.2010.	8E	A4	WAUZZZ8E13A...	AUDI	2003	1450	M1	1	195/65R15-235/...	74	1896	D
NULL	14.1.2010.	8E	A4	WAUZZZ8E13A...	AUDI	2003	1525	M1	1	195/65R15 91V,...	96	1896	D
NULL	7.1.2010.	8E	A4	WAUZZZ8E22A...	AUDI	2002	1525	M1	1	195/65R15-235/...	96	1896	D
NULL	4.1.2010.	8E	A4	WAUZZZ8E55A...	AUDI	2005	1690	M1	1	235/45 R 17 9...	188	3123	B
NULL	7.1.2010.	8E	A4	WAUZZZ8E56A...	AUDI	2006	1565	M1	1	205/55R16 - 235...	103	1968	D
NULL	13.1.2010.	8E	A4	WAUZZZ8E72A...	AUDI	2002	1525	M1	1	195/65R15 91V,...	96	1896	D
NULL	19.1.2010.	8L	A3	WAUZZZ8L62A...	AUDI	2002	1260	M1	1	195/65 R15 91H	74	1896	D
NULL	22.1.2010.	8L	A3	WAUZZZ8LZ1A...	AUDI	2001	1165	M1	1	195/65R15 91V,...	75	1596	B
NULL	4.1.2010.	8P	A3	WAUZZZ8P14A...	AUDI	2004	1355	M1	1	205/55R16 91H...	75	1595	B
NULL	13.1.2010.	8P	A3	WAUZZZ8P1AA...	AUDI	2010	1090	M1	1	225/45R17 91W;...	77	1598	D
NULL	25.1.2010.	8P	A3	WAUZZZ8P24A...	AUDI	2004	1415	M1	1	205/55 R16 91W	103	1968	D
NULL	25.1.2010.	8R	Q5	WAUZZZ8R3AA...	AUDI	2010	1820	M1	1	235/55 R19 101V	125	1968	D
NULL	15.1.2010.	8Z	A2	WAUZZZ8ZZ1N...	AUDI	2001	1065	M1	1	155/65R15 - 195...	55	1422	D
NULL	13.1.2010.	346	320	WBAAL71000K...	BAYERISCHE M...	2000	1450	M1	1	195/65R15 91H,...	100	1951	D
NULL	19.1.2010.	346	320	WBAAL71020C...	BAYERISCHE M...	2000	1450	M1	1	195/65R15-225/...	100	1951	D

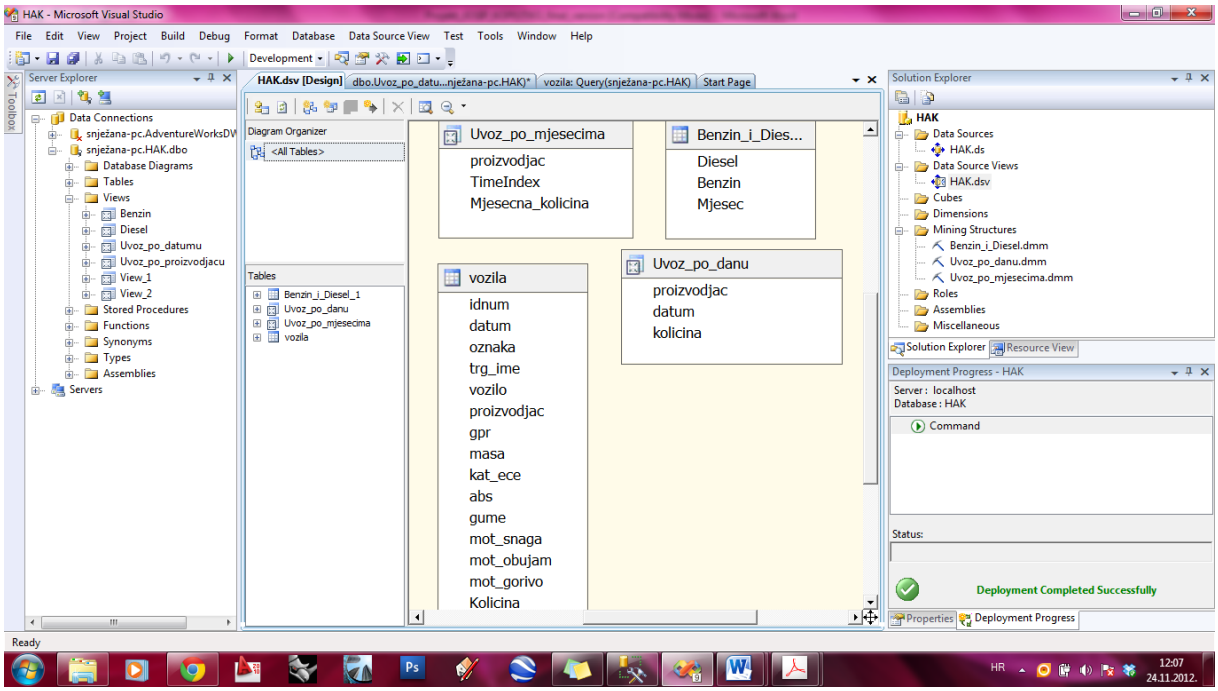
1 of 17458

Slika 4-2. HAK baza sa stvarnim podacima

4.2 Otkrivanje znanja pomoću Microsoft Time Series algoritma

Kao što je prije rečeno, strukturiranjem baze podataka dolazi se do važnih informacija. Sljedeći korak u BI procesu jest da se te podatke pretvori u korisna znanja kao pomoć pri donošenju ključnih poslovnih odluka. [3],[4]

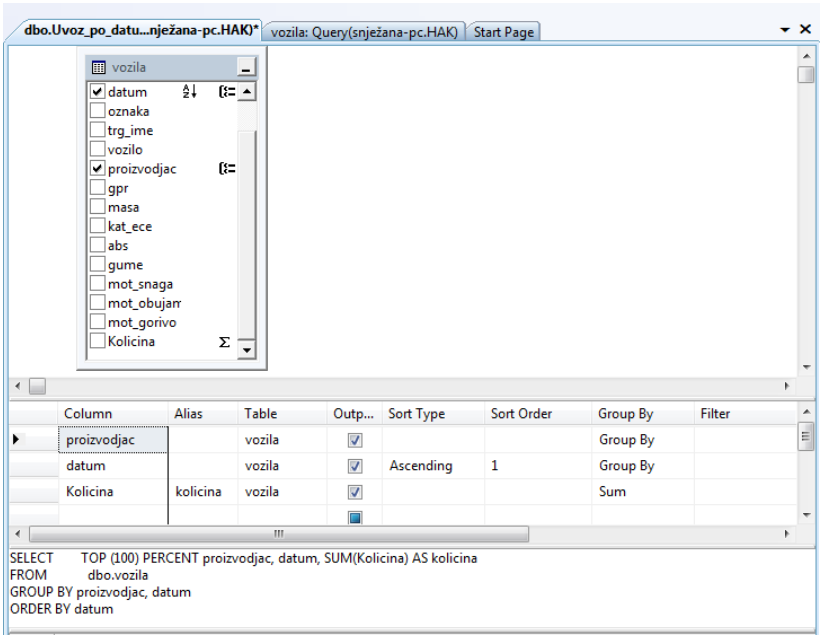
Algoritam vremenskih nizova (eng. Microsoft Time Series algorithm) kako je navedeno u prijašnjim poglavljima, koristi se za analizu i predviđanje vremenski ovisnih podataka. Ujedno, on je kombinacija sljedeća dva algoritma: ARTxp (AutoRegressionTree) i ARIMA (Auto-Regressive Integrated Moving Average) algoritma. ARTxp algoritam se najčešće koristi kod kratkoročnih predviđanja dok je ARIMA algoritam bolji za dugoročna predviđanja. Microsoft Time Series algoritam, po programski zadanim postavkama, stapa rezultate oba algoritma kako bi što bolje odredio kako kratkoročna, tako i dugoročna predviđanja. U sljedećim poglavljima pokušat će se primijeniti ovaj algoritam na dva različita seta podataka s različito definiranim vremenskim nizovima kako bi se iz ove baze podataka moglo izvući što više korisnih znanja. Na slici 4-3 prikazan je pogled izvora podataka sa potrebnim tablicama koje ćemo koristiti za rudarenje podataka. Kako su dobivene, bit će objašnjeno u sljedećim konkretnim primjerima.



Slika 4-3. Pogled izvora podataka u SASS-u HAK

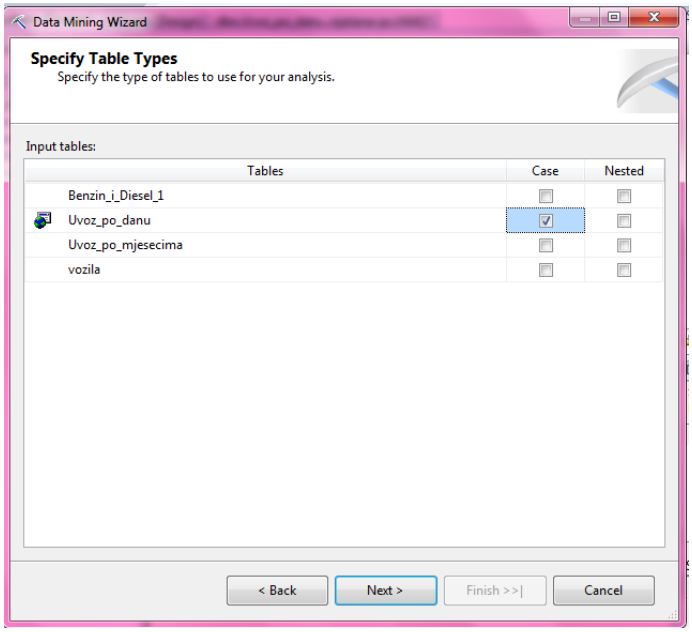
4.2.1 Predviđanje uvoza – budući trendovi uvoza rabljenih automobila

Kako bi se dao konkretniji odgovor na pitanje kakav će biti budući trend uvoza rabljenih automobila, dobiveni podaci podvrgnut će se analizi pomoću Microsoft Time Series algoritma. Za ovu vrstu algoritma nije izričito potrebno raditi posebne upite na bazu podataka no kako bi se olakšao pregled i kako bi model radio samo sa potrebnim atributima i zapisima napravit će se upit koji će nam iz baze podataka izvući sljedeće atribute potrebne za analizu: datum, proizvodjac, količina. Ovaj upit/pogled prikazan je na slici 4-3 pod imenom uvoz_po_datumu.



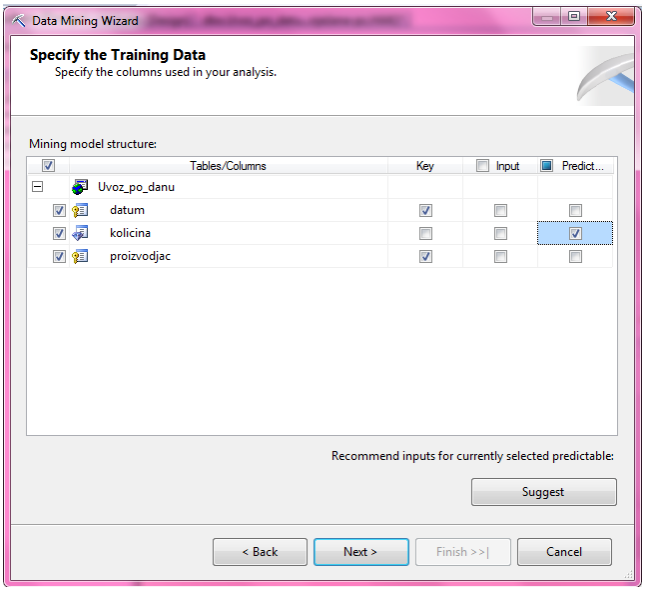
Slika 4-4. Prikaz upita uvoz_po_datumu

Sljedeći korak je ubacivanje ovog pogleda u SSAS što je i učinjeno na slici 4-3. Ubacivanjem u SSAS ti podaci postaju upotrebljivi za analizu rudarenjem podataka. Lijevim klikom miša na folder Mining Structures u Solution Explorer prozoru odabiremo opciju New Mining Structure u koji ćemo dodati novi model rudarenja podataka vezan uz prethodno ubačenu tablicu u Data Source View. U Data Mining Wizardu (slika 4-5) odaberemo tablicu nad kojom želimo izvršiti rudarenje.

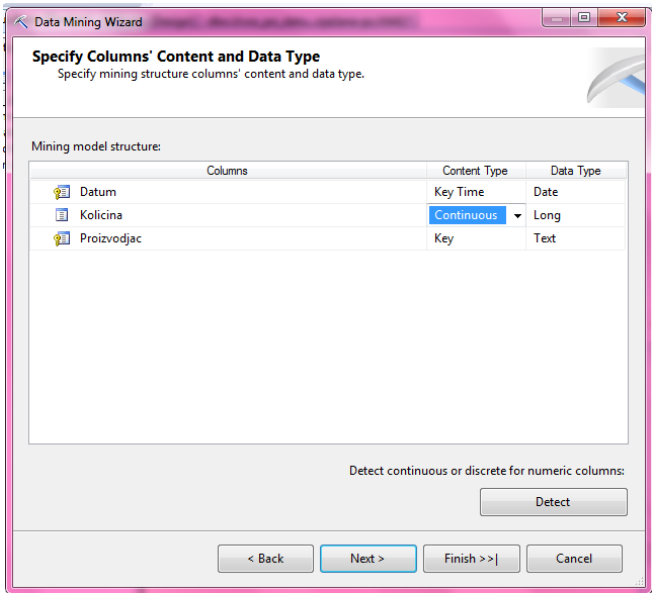


Slika 4-5. Odabir tablice uvoz_po_danu za rudarenje

Sljedeće što slijedi jest odabrati ulazne i izlazne odnosno predvidljive kolone prikazane na slici 4-6. Ne smije se smetnuti s uma da su predvidljivi podaci ujedno i ulazni. U ovom slučaju predviđamo količinu pa će biti definirana kao podatak koji želimo predvidjeti. Kao Key kolone definirani su datum i proizvodjac pošto želimo predvidjeti koje će se vrste automobila i u kojoj količini uvažati u određenom vremenskom periodu. Na slici 4-7 prikazani su sadržaji i vrste podataka. Predvidljivi podaci za ovu vrstu algoritma uvijek moraju biti kontinuiranog sadržaja.

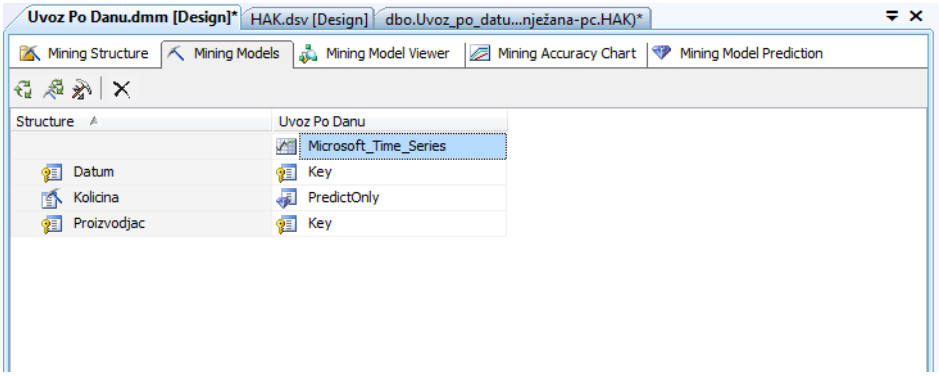


Slika 4-6. Označavanje predvidljivih kolona za uvoz_po_danu



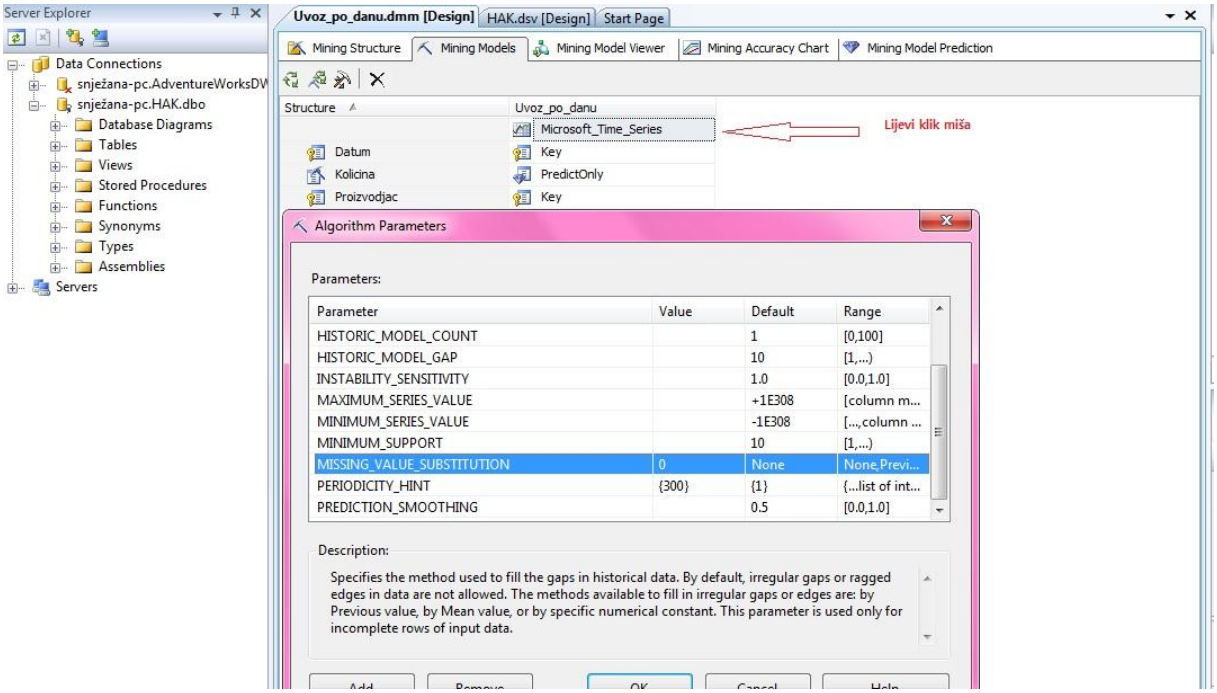
Slika 4-7. Odabir sadržaja i vrste podataka za uvoz_po_danu

Na slici 4-8 prikazan je izgled Mining Models tab-a na kojem je moguće vidjeti prethodno definirani model nazvan Uvoz Po Danu.



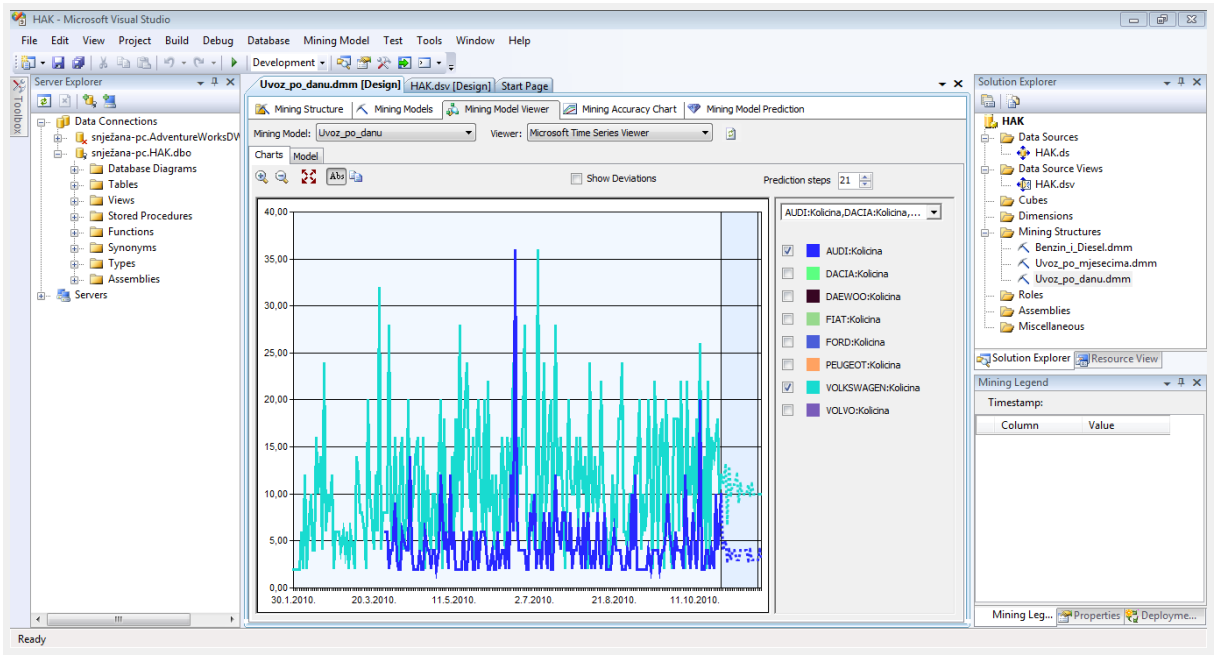
Slika 4-8. Prikaz Mining Models tab-a Uvoz Po Danu

Kako ovaj algoritam radi isključivo sa vremenski kontinuiranim predvidljivim varijablama potrebno je podesiti parametre algoritma na način prikazan slikom 4-9 zato što se događa da u nekom vremenskom periodu nije uvežena neka od vrsti automobila pa tu ujedno nastaje i praznina u vremenskoj krivulji. Da bi se taj problem riješio potrebno je pod MISSING_VALUE_SUBSTITUTION upisati vrijednost 0, što bi značilo da tamo gdje postoji praznina jednostavno nije bilo uveženih automobila. Pošto smo podatke ubacivali na dnevnoj bazi PERIODICITY_HINT podešen je na {300}.

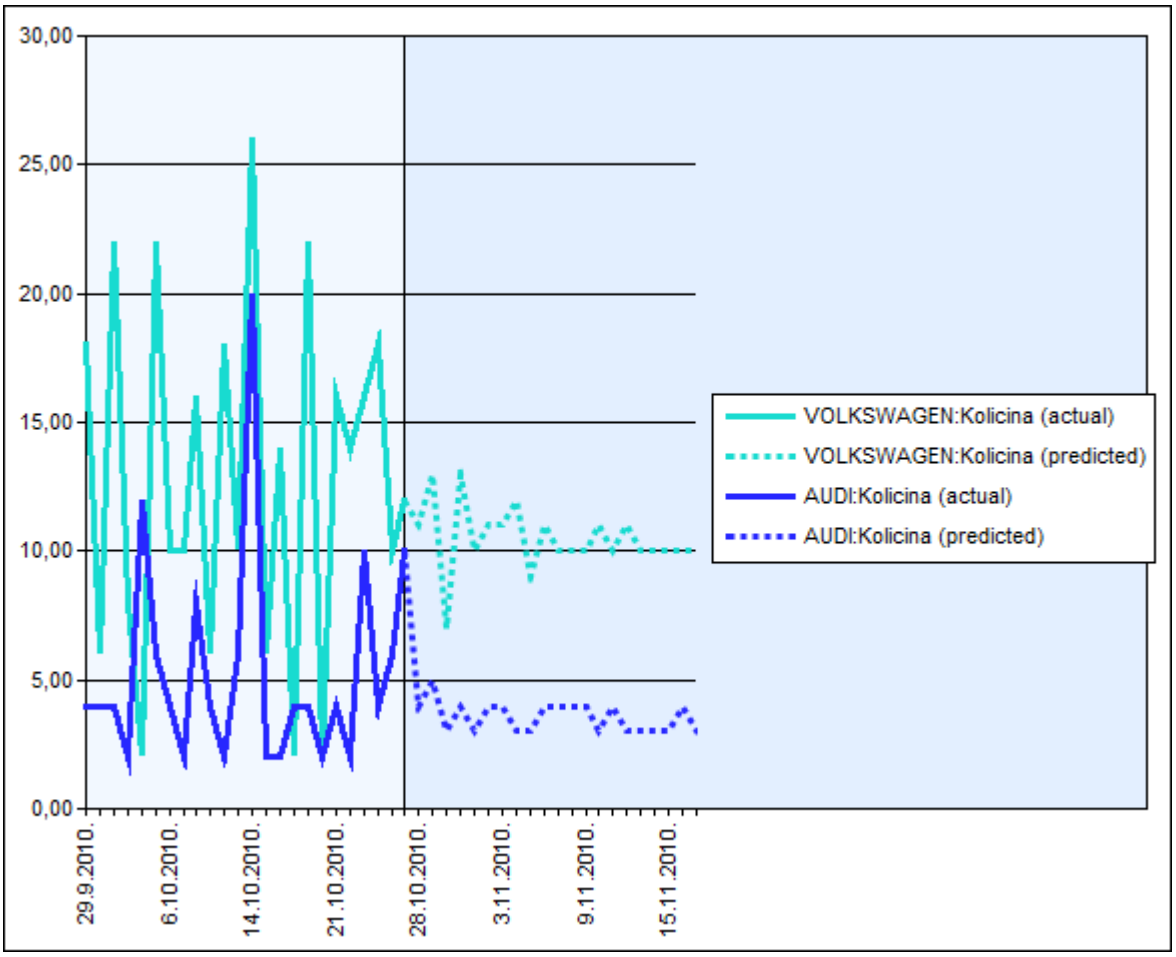


Slika 4-9. Prikaz konfiguracije parametara za Uvoz po Danu

Sljedeće što preostaje jest prikazati rezultate dobivene primjenom Microsoft Time Series algoritomom. Rezultati su prikazani na slici 4-10 gdje su u Chart prikazu na drop down listi izabrane vrste automobila za čije količine želimo obaviti predviđanje, točnije, za koje želimo vidjeti kako će se u budućnosti kretati trend uvoza odabranih modela. Od nekolicine odabranih modela na slici 4-11 su prikazane krivulje za Audi i Volkswagen.

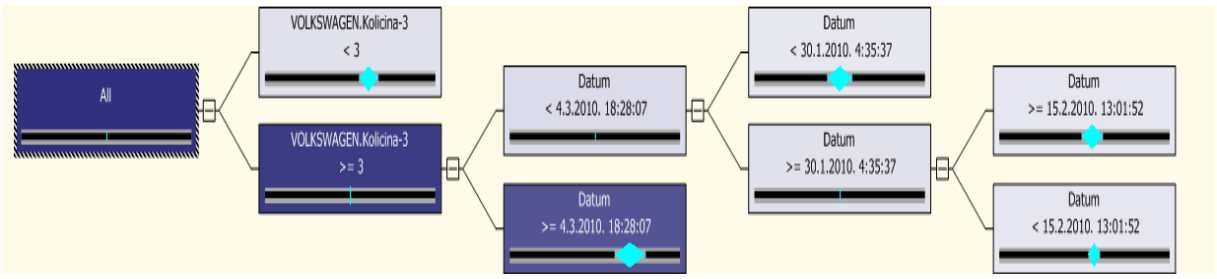


Slika 4-10. Prikaz rezultata Microsoft Time Series algoritma

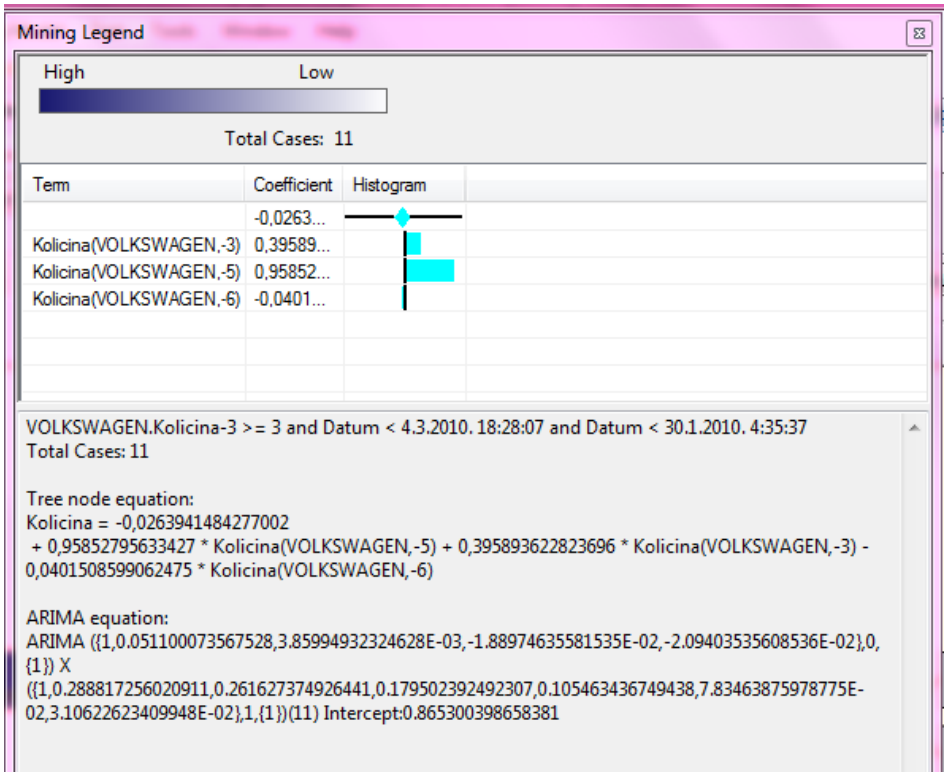


Slika 4-11. Prikaz predviđanja količina - Audi i Volkswagen

Sljedeća slika 4-12 prikazuje proces dobivanja rezultata u Model prikazu za Volkswagen marku automobila. Mining Legend prozor (slika 4-13) pokazuje informacije za ovaj algoritam, i to: koeficijente, histograme i jednadžbe čvorova stabla. Pristup tim informacijama je omogućen kako bi se bolje razumjela metoda pomoću koje su napravljena predviđanja.



Slika 4-12. Model prikaz za Volkswagen

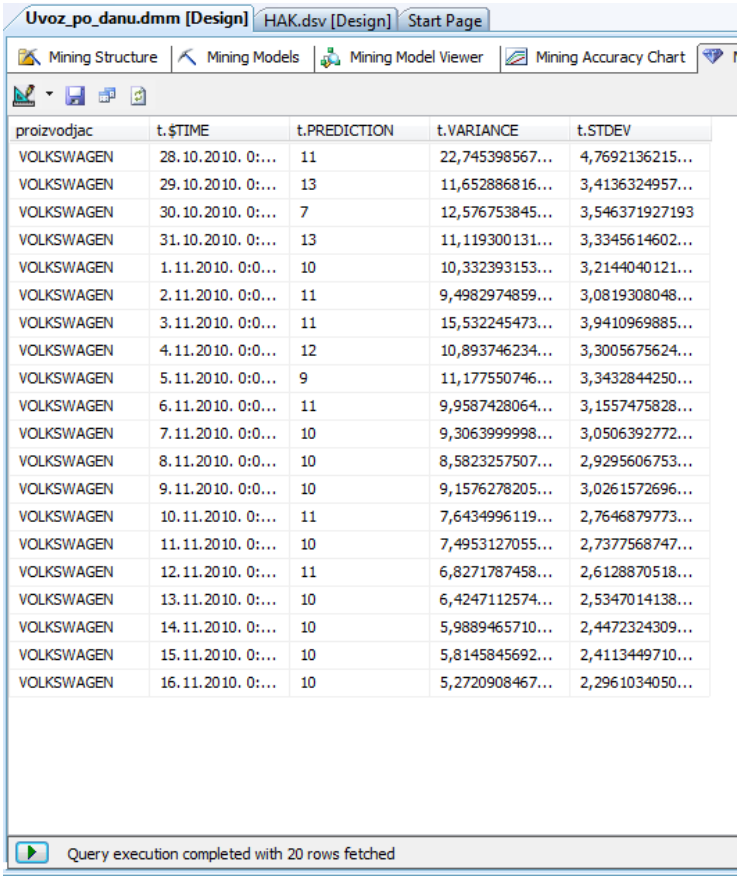


Slika 4-13. Mining Legend prozor sa jednadžbama

Slijedi pisanje DMX (Data Mining Exstension) upita u Mining Model Prediction prikazu. DMX upit će nam prikazati predviđene vrijednosti za odabrani vremenski period koji se kasnije mogu spremiti kao zasebna tablica. Rezultati DMX upita su prikazani na slici 4-14.

DMX upit glasi:

```
SELECT FLATTENED [proizvodjac],
(SELECT
    $Time,
    [kolicina] as [PREDICTION],
    PredictVariance([kolicina]) AS [VARIANCE],
    PredictStdev([kolicina]) AS [STDEV]
FROM
     PredictTimeSeries([kolicina], 20) AS t
) AS t
FROM Uvoz_po_danu
WHERE [proizvodjac] = 'VOLKSWAGEN'
```



proizvodjac	t.$TIME	t.PREDICTION	t.VARIANCE	t.STDEV
VOLKSWAGEN	28.10.2010. 0:...	11	22,745398567...	4,7692136215...
VOLKSWAGEN	29.10.2010. 0:...	13	11,652886816...	3,4136324957...
VOLKSWAGEN	30.10.2010. 0:...	7	12,576753845...	3,546371927193
VOLKSWAGEN	31.10.2010. 0:...	13	11,119300131...	3,3345614602...
VOLKSWAGEN	1.11.2010. 0:0...	10	10,332393153...	3,2144040121...
VOLKSWAGEN	2.11.2010. 0:0...	11	9,4982974859...	3,0819308048...
VOLKSWAGEN	3.11.2010. 0:0...	11	15,532245473...	3,9410969885...
VOLKSWAGEN	4.11.2010. 0:0...	12	10,893746234...	3,3005675624...
VOLKSWAGEN	5.11.2010. 0:0...	9	11,177550746...	3,3432844250...
VOLKSWAGEN	6.11.2010. 0:0...	11	9,9587428064...	3,1557475828...
VOLKSWAGEN	7.11.2010. 0:0...	10	9,3063999998...	3,0506392772...
VOLKSWAGEN	8.11.2010. 0:0...	10	8,5823257507...	2,9295606753...
VOLKSWAGEN	9.11.2010. 0:0...	10	9,1576278205...	3,0261572696...
VOLKSWAGEN	10.11.2010. 0:...	11	7,6434996119...	2,7646879773...
VOLKSWAGEN	11.11.2010. 0:...	10	7,4953127055...	2,7377568747...
VOLKSWAGEN	12.11.2010. 0:...	11	6,8271787458...	2,6128870518...
VOLKSWAGEN	13.11.2010. 0:...	10	6,4247112574...	2,5347014138...
VOLKSWAGEN	14.11.2010. 0:...	10	5,9889465710...	2,4472324309...
VOLKSWAGEN	15.11.2010. 0:...	10	5,8145845692...	2,4113449710...
VOLKSWAGEN	16.11.2010. 0:...	10	5,2720908467...	2,2961034050...

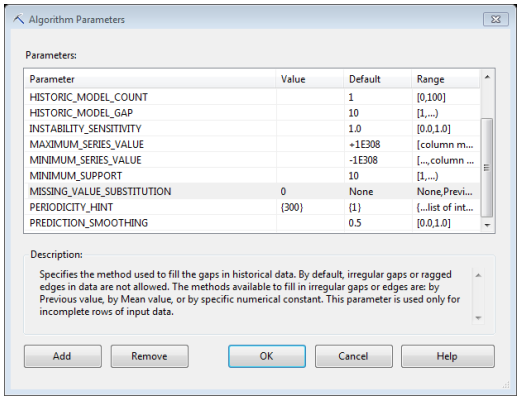
Slika 4-14. Rezultati DMX upita za Uvoz_po_danu

Sada se na ove podatke mogu primjeniti daljnje statističke obrade kako bi se dobili odgovarajući parametri točnosti ovog modela.

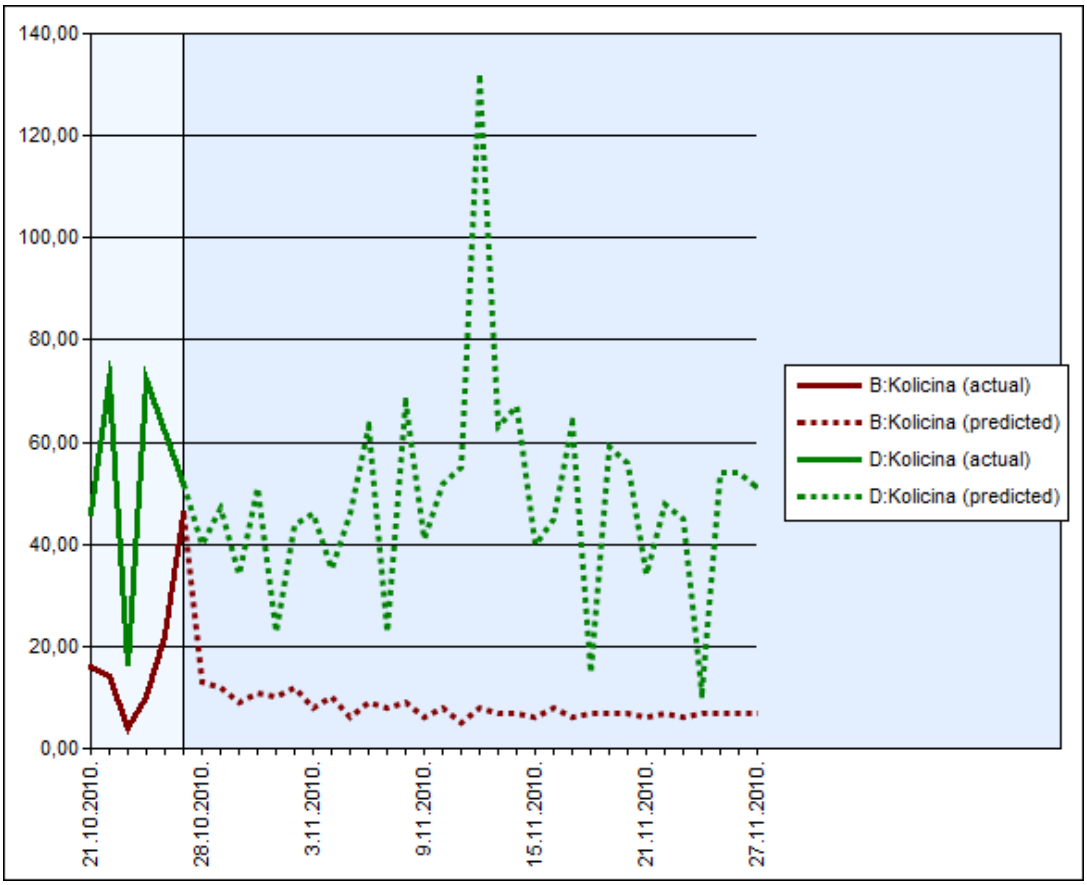
U konačnici, zašto su ovo važni podaci? Treba uzeti u obzir dinamiku današnjeg tržišta te socijalno stanje prosječnog građanina Hrvatske uz pretpostavku da suvremeni način života nameće automobil kao nužnost. Naime, prognoze s tržišta navješćuju da će se dramatičan trend uvoza i prodaje (godišnji pad od 17-18%) novih automobila nastaviti i vjerojatno tako skoro psihološka granica od dvije tisuće novo-registriranih automobila mjesečno neće biti prijeđena. Zadnje dvije godine tržište rabljenih automobila je u laganom porastu od 7-8%, što bi značilo da će se i prosječna starost automobila na hrvatskim cestama povisiti. Ova informacija bi mogla pomoći i predložiti auto-salonima da počnu pojačano oglašavati svoj program rabljenih vozila i servisa računajući na povećanu opreznost kupaca. Dobavljačima zamjenskih dijelova moglo bi biti zanimljiva u vidu predviđanja i planiranja svojih zaliha na skladištu.

4.2.2 Uvoz rabljenih automobila ovisno o motornom gorivu

Ovaj model isto tako koristi Microsoft Time Series algoritam kako bi predvidio buduće kretanje uvoza automobila ovisno o vrsti goriva. Model je koncipiran na način da su prikazane dvije krivulje, jedna za vozila na dizel motorna goriva, a druga na benzin. Vremenski periodi prikazani su na dnevnoj bazi pa su potrebna određena podešenja algoritamskih parametara. Na slici 4-15 su prikazani algoritamski parametri, a na slici 4-16 rezultati rudarenja podataka.



Slika 4-15. Konfiguracija parametara za predviđanje vrste motornih goriva



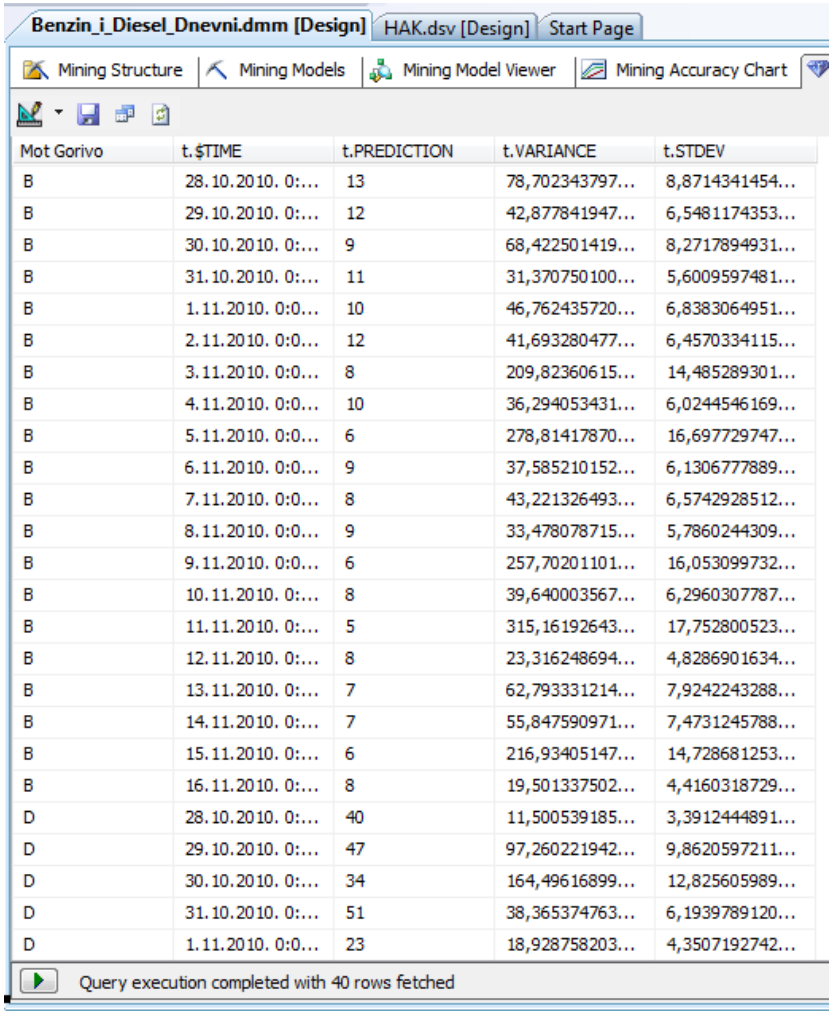
Slika 4-16. Rezultati predviđanja uvoza ovisno o motornom gorivu

Sada slijedi pisanje DMX upita kako bi imali preglednije podatke o predviđanju. U DMX upitu će nam biti prikazane vrijednosti predviđanja za narednih 20 dana za obje vrste goriva.

DMX upit glasi:

```
SELECT FLATTENED [Mot Gorivo],
(SELECT
    $Time,
    [Kolicina] as [PREDICTION],
    PredictVariance([Kolicina]) AS [VARIANCE],
    PredictStdev([Kolicina]) AS [STDEV]
FROM
     PredictTimeSeries([Kolicina], 20) AS t
) AS t
FROM Benzin_i_Diesel_Dnevni
WHERE [Mot Gorivo] = 'D' OR [Mot Gorivo] = 'B'
```

Rezultati DMX upita prikazani su na slici 4-17 i služe za daljnu statističku obradu.

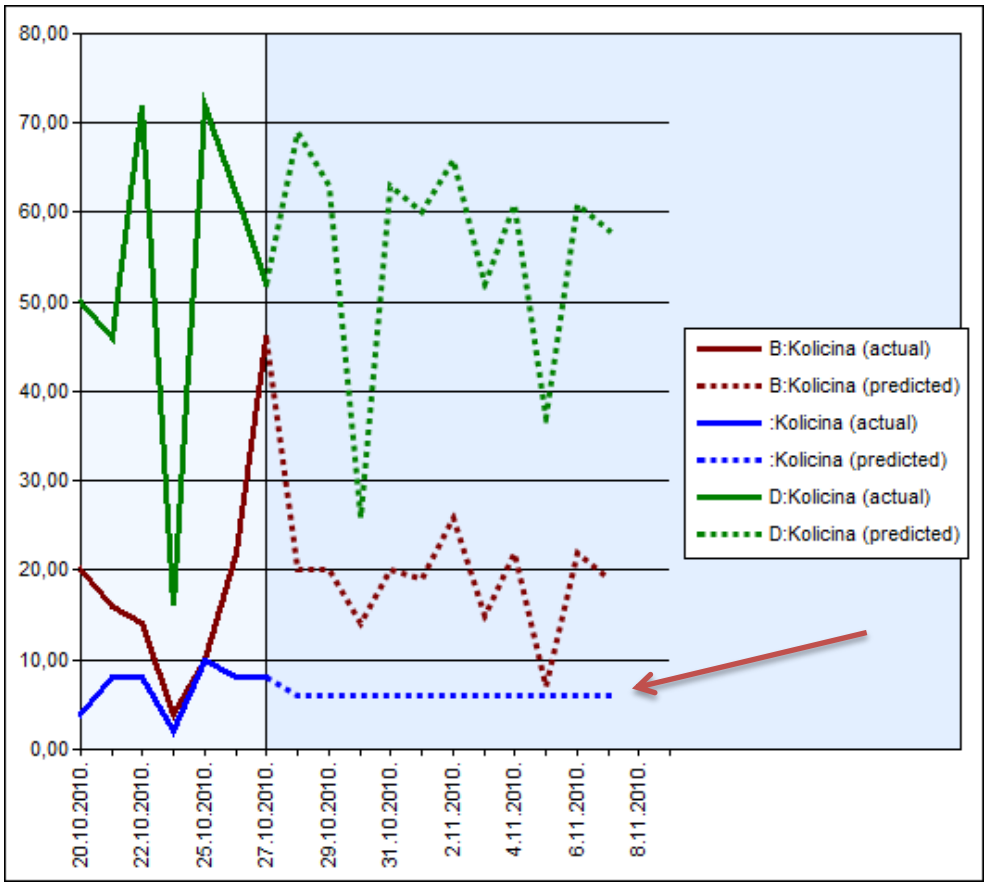


Mot Gorivo	t.$TIME	t.PREDICTION	t.VARIANCE	t.STDEV
B	28.10.2010. 0:...	13	78,702343797...	8,8714341454...
B	29.10.2010. 0:...	12	42,877841947...	6,5481174353...
B	30.10.2010. 0:...	9	68,422501419...	8,2717894931...
B	31.10.2010. 0:...	11	31,370750100...	5,6009597481...
B	1.11.2010. 0:0...	10	46,762435720...	6,8383064951...
B	2.11.2010. 0:0...	12	41,693280477...	6,4570334115...
B	3.11.2010. 0:0...	8	209,82360615...	14,485289301...
B	4.11.2010. 0:0...	10	36,294053431...	6,0244546169...
B	5.11.2010. 0:0...	6	278,81417870...	16,697729747...
B	6.11.2010. 0:0...	9	37,585210152...	6,1306777889...
B	7.11.2010. 0:0...	8	43,221326493...	6,5742928512...
B	8.11.2010. 0:0...	9	33,478078715...	5,7860244309...
B	9.11.2010. 0:0...	6	257,70201101...	16,053099732...
B	10.11.2010. 0:...	8	39,640003567...	6,2960307787...
B	11.11.2010. 0:...	5	315,16192643...	17,752800523...
B	12.11.2010. 0:...	8	23,316248694...	4,8286901634...
B	13.11.2010. 0:...	7	62,793331214...	7,9242243288...
B	14.11.2010. 0:...	7	55,847590971...	7,4731245788...
B	15.11.2010. 0:...	6	216,93405147...	14,728681253...
B	16.11.2010. 0:...	8	19,501337502...	4,4160318729...
D	28.10.2010. 0:...	40	11,500539185...	3,3912444891...
D	29.10.2010. 0:...	47	97,260221942...	9,8620597211...
D	30.10.2010. 0:...	34	164,49616899...	12,825605989...
D	31.10.2010. 0:...	51	38,365374763...	6,1939789120...
D	1.11.2010. 0:0...	23	18,928758203...	4,3507192742...

Slika 4-17. Rezultati DMX upita za vrstu motornog goriva

4.2.3 Cross-prediction metoda na primjeru uvoza automobila prema motornom gorivu

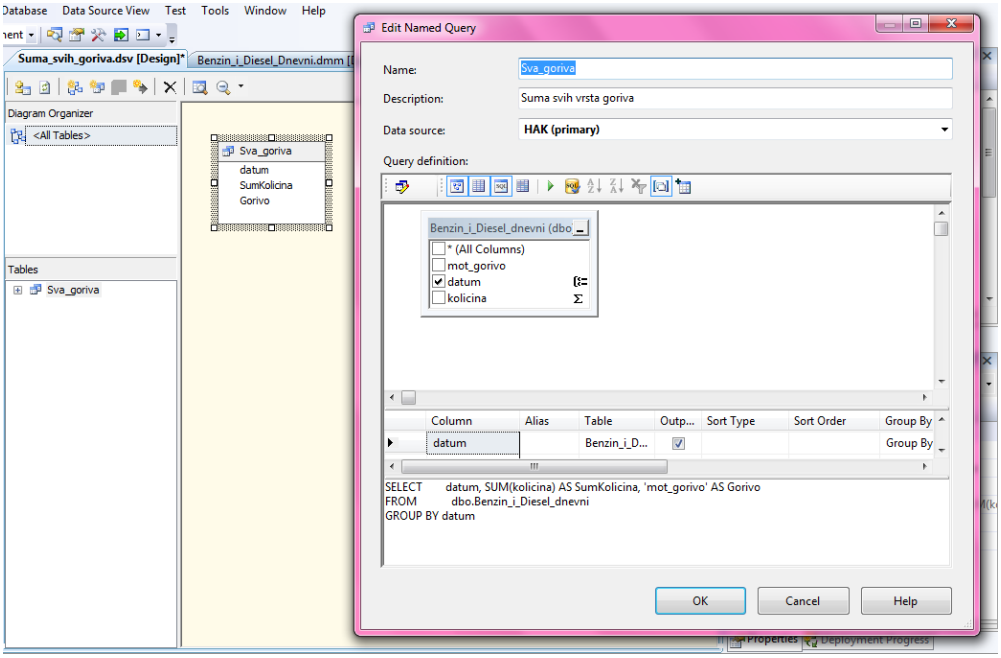
U ovom će se primjeru obraditi cross-prediction metoda te objasniti njena važnost kod poboljšanja predviđanja budućih vrijednosti. Cross prediction metoda se koristi za poboljšanje područja u grafu za koja se ne dobivaju ili izostaju rezultati primjenom Microsoft Time Series algoritma. Jedan takav graf prikazan je slikom 4-18. Iz grafa se mogu očitati vrijednosti predviđanja za dizelska i benzinska motorna goriva dok za slučaj ostalih goriva (plin, električna energija...), prikazanih plavom isprekidanom linijom, rezultati izostaju.



Slika 4-18. Predviđanje uvoza ovisno u vrsti goriva

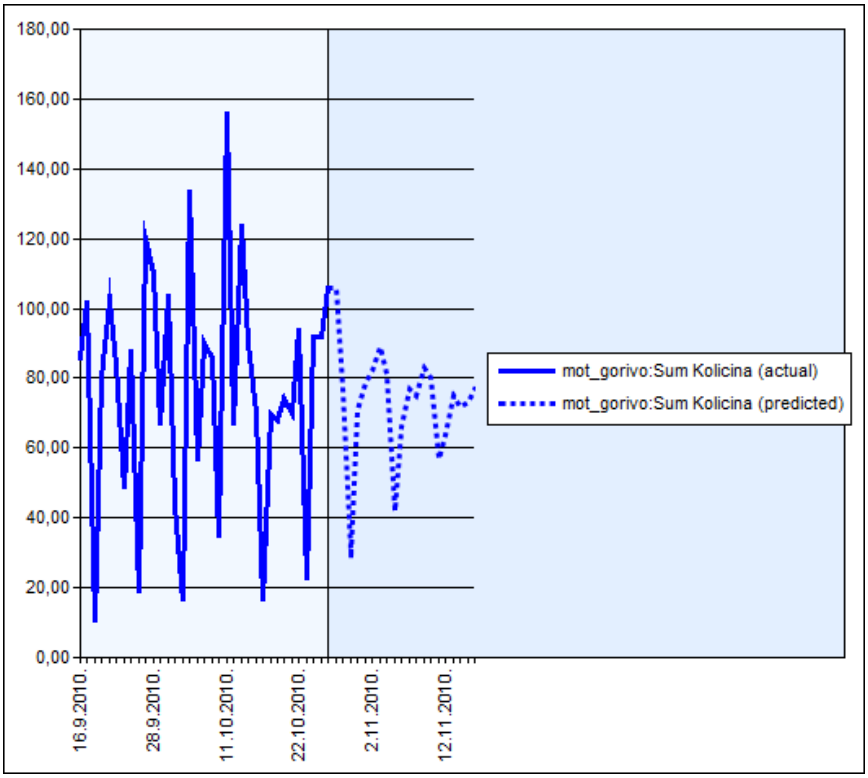
Cross- prediction metoda pomoći će u boljem generiranju predviđajućih vrijednosti za ostale vrste goriva. Metoda se temelji na predviđanju generalnih trendova sa boljim predviđanjima i njihovom ″ukrižavanju″ sa trendovima koji daju slabija, stagnirajuća predviđanja. U nastavku će biti objašnjeni koraci za provođenje cross-prediction metode.

Prvo treba napraviti generalni model pod nazivom ″Sva_goriva″ za koji smo zaključili da dobro predviđa buduća događanja. Upit za izradu generalnog modela i njegovog pripadajućeg Data Source View-a prikazan je na slici 4-19. Slika 4-20 prikazuje rezultate data mining-a za generalni model.



Slika 4-19. Upit za izradu generalnog modela

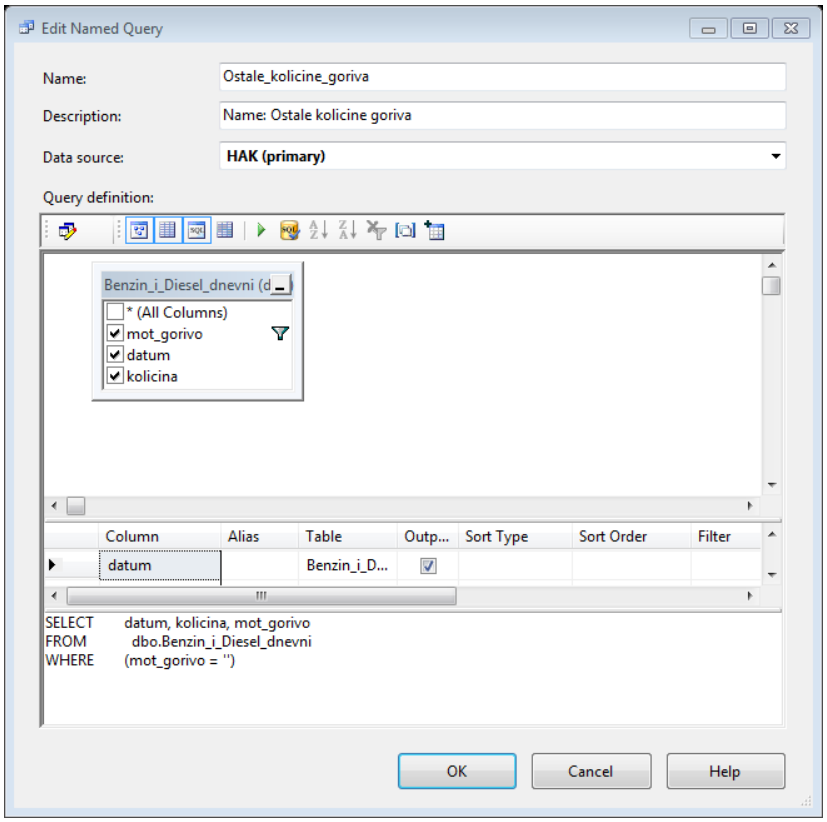
Nakon toga preostaje prikazati rezultate tog upita koji su prikazani slikom 4-20.



Slika 4-20. Rezultati generalnog modela

Sljedeće preostaje napraviti model na kojem će se pomoću cross-prediction metode poboljšati rezultati predviđanja, tj. model ostalih vrsta goriva za koje je prethodno zaključeno

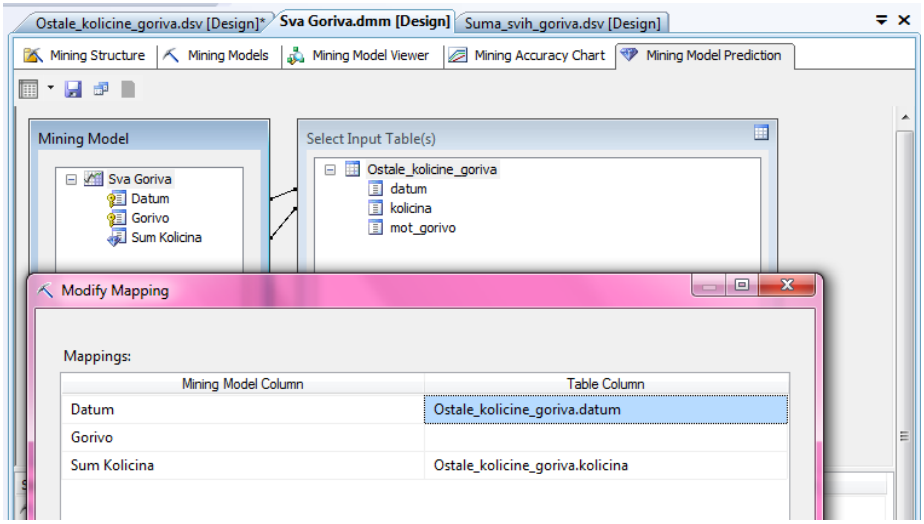
da rezultati predviđanja izostaju. Slikom 4-21 prikazan je upit za izradu modela pod nazivom ″Ostale_kolicine_goriva″ na kojem će biti izvršen cross-prediction.



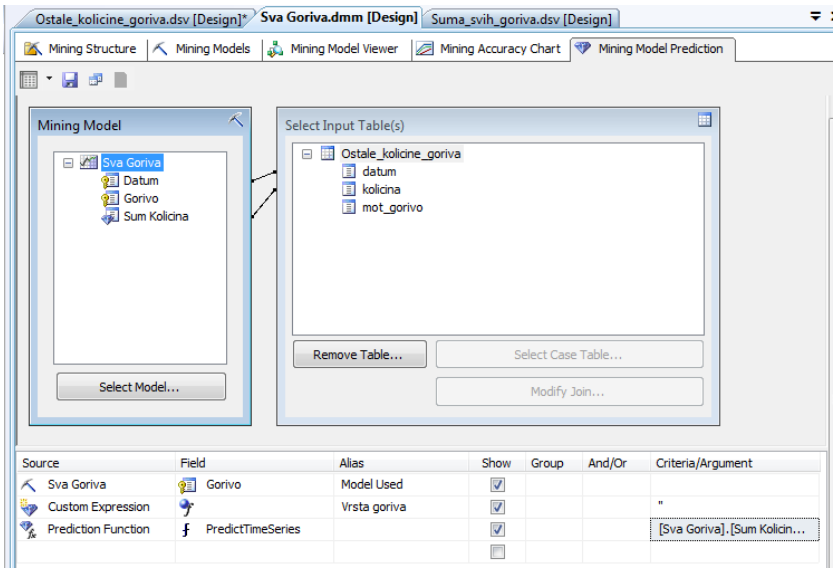
Slika 4-21. Upit za izradu modela ostalih goriva

Slika 4-21 prikazuje upit za izradu modela ostalih goriva koji su bazi podataka navedeni pod atributom mot_gorivo i prikazani praznom ćelijom, a odonose se na ostala goriva.

Nakon dobivenih rezultata generalnog modela koji je definiran kao zbroj svih vrsta goriva i modela kojem će se poboljšavati preformanse predviđanja može se krenuti na primjenu generalnog modela kao referentog modela za predviđanje budućih vrijednosti ostalih goriva (plin, električna energija...). Slika 4-22 prikazuje odabir i prilagođavanje veza između tablica za provođenje cross-prediction metode, dok slika 4-23 prikazuje ″Design″ prikaz za provedbu cross-prediction metode napravljen u Mining Model Prediction tab-u. ″Design″ prikaz je user-friendly prikaz za kojeg korisnik ne mora razumijeti naredbe za kreiranje upita.



Slika 4-22. Odabir i prilagođavanje veza između cross-prediction tablica



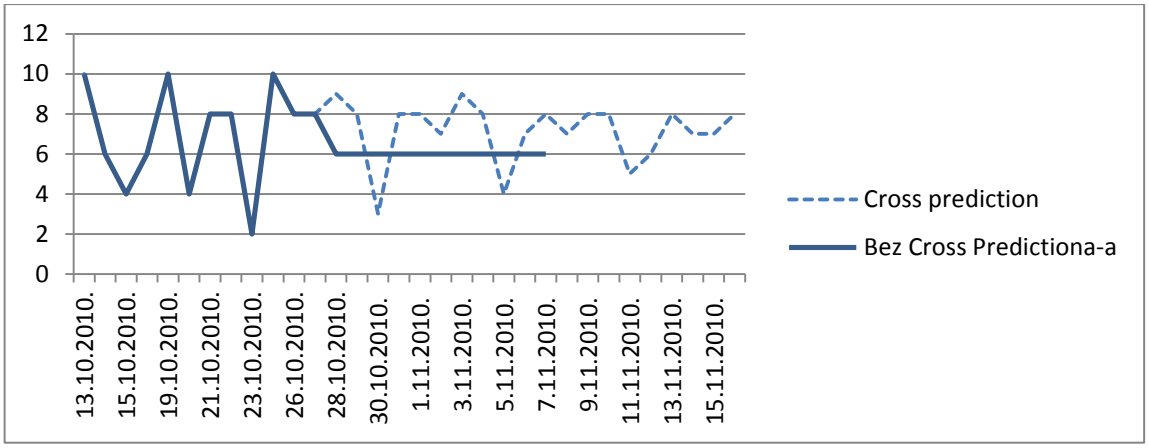
Slika 4-23. Design prikaz za cross-prediction metodu

Upit kreiran u ″Query″ prikazu izgleda na sljedeći način:

```
SELECT
  ([Sva Goriva].[Gorivo]) as [Model Used],
  ( '') as [Vrsta goriva],
  PredictTimeSeries([Sva Goriva].[Sum Kolicina],20,REPLACE_MODEL_CASES)
From
  [Sva Goriva]
PREDICTION JOIN
  OPENQUERY([HAK],
    'SELECT
      [datum],
      [kolicina]
    FROM
      (SELECT       datum, kolicina, mot_gorivo
```

```
FROM          dbo.Benzin_i_Diesel_dnevni
WHERE       (mot_gorivo = "")) as [Ostale_kolicine_goriva]
   ') AS t
ON
  [Sva Goriva].[Datum] = t.[datum] AND
       [Sva Goriva].[Sum Kolicina] = t.[kolicina]
```

Slika 4-24 prikazuje rezultate cross-prediction metode upotrebljene za poboljšanje predviđanja vrijednosti uvoza automobila pogonjenih na ostala goriva (plin, električna energija...).



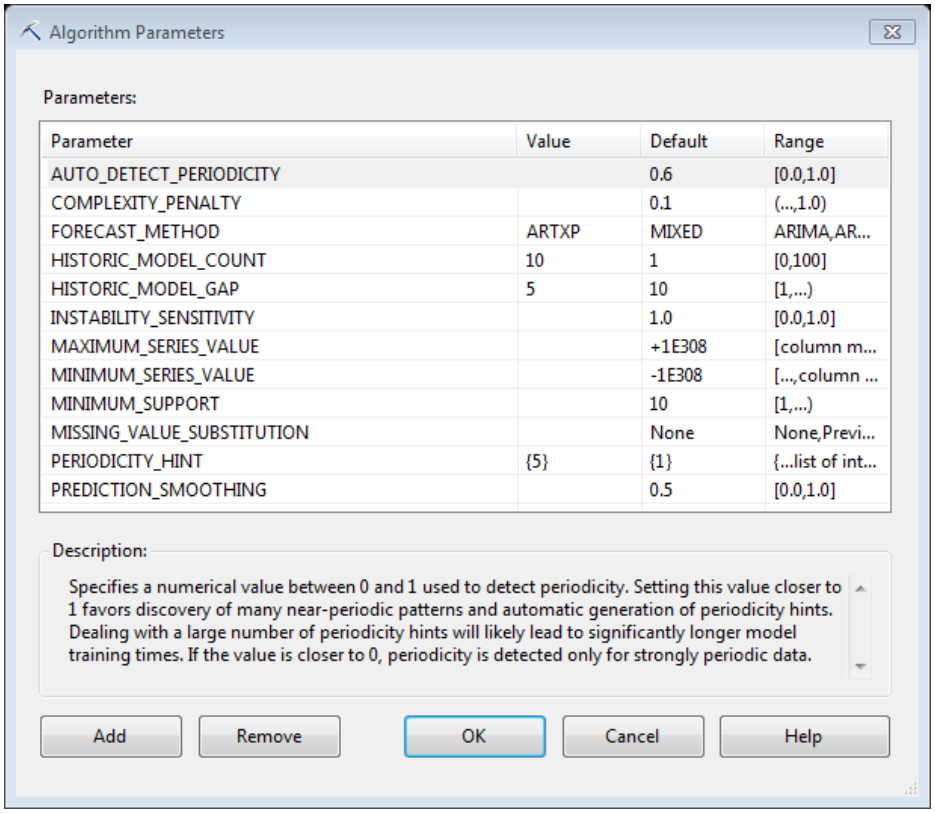
Slika 4-24. Rezultat Cross prediction metode

5. Validacija modela

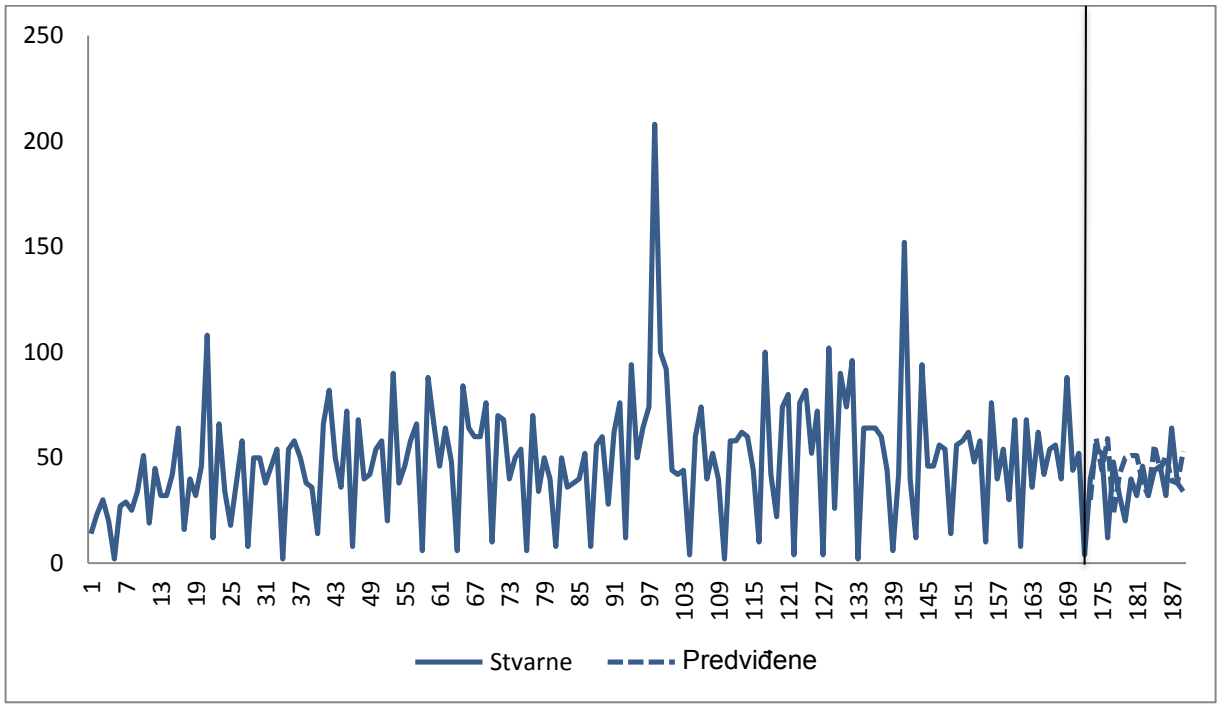
U ovom poglavlju pokušat će se provjeriti na kojoj razini predviđanja algoritam radi, tj. koliko dobro predviđa buduće događaje ili vrijednosti i to na primjeru predviđanja uvoza rabljenih automobila s Diesel agregatima.

5.1 Validacija modela predviđanja rezultata na dnevnoj bazi

U svrhu toga napravljen je novi set podataka koji sadrži stvarne rezultate prvih sedam mjeseci na kojima će se ujedno i trenirati model. Konfiguracijom parametara (slika 5-1) te provedbom algoritma nad novom tablicom dobiveni su predviđeni rezultati prikazani na slici 5-2, a odnose se na automobile s Diesel agregatom s predviđanjima u sljedećih 17 dana.

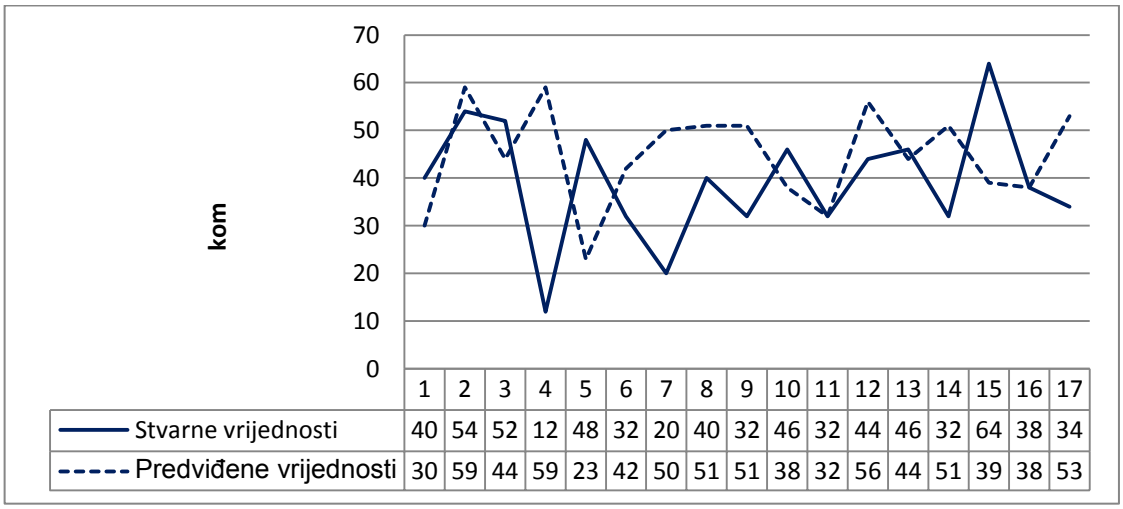


Slika 5-1. Konfiguracija parametara



Slika 5-2. Rezultati predviđanja

Na slici 5-3 su pobliže prikazane predviđene i stvarne vrijednosti.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Stvarne vrijednosti	40	54	52	12	48	32	20	40	32	46	32	44	46	32	64	38	34
Predviđene vrijednosti	30	59	44	59	23	42	50	51	51	38	32	56	44	51	39	38	53

Slika 5-3. Graf stvarnih i predviđenih vrijednosti

Sljedeće što slijedi jest provesti statističku analizu nad stvarnim i predviđenim podacima kako bi se dobila informacija koliko dobro ovaj model predviđa buduće događaje/vrijednosti.

Daljnjom statističkom obradom ustanovit će se je li postoji značajna razlika između predviđenih i stvarnih vrijednosti. Za statističku analizu korištena je korelacija i T-test

smješteni u *Microsoft Excel Data Analysis Tool*-u. Rezultati analize prikazani su tablicom 5-2.

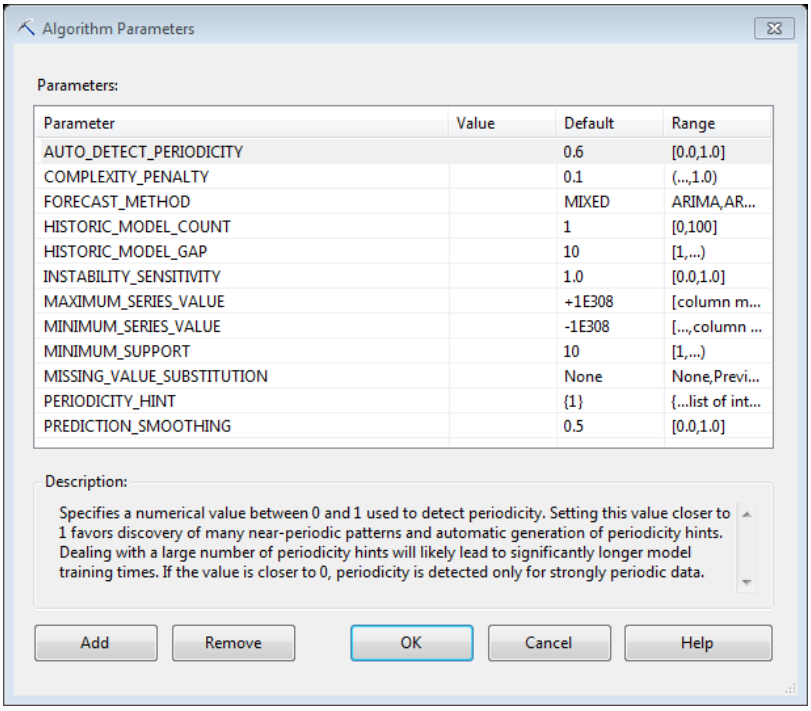
Tablica 5-1. Rezultati analize

	Stupanj korelacije r	***P vrijednost t-testa (α=0,05)***
Stvarne vrijednosti	- 0,304	0,172
Predviđene vrijednosti		

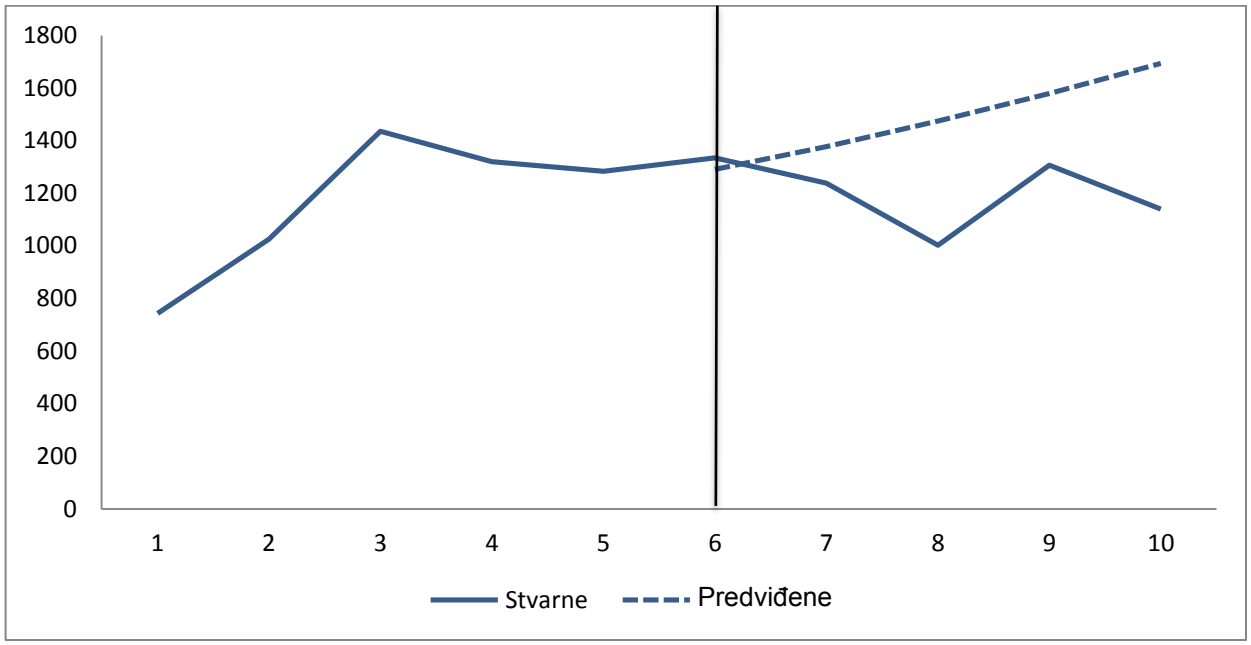
Iz tablice 5-2 je vidljiva P vrijednost koja iznosi 0,172. Kako je P vrijednost veća od 0,05, prihvaćamo nultu hipotezu koja glasi da se stvarni i predviđajući podaci značajno ne razlikuju te se na osnovu T-testa moze zaključuje da nema značajne razlike. Prema koeficijentu korelacije može se zaključiti da predviđene vrijednosti ne prate stvarne vrijednosti. U konačnici, ovaj model zbog male i negativne korelacije ne predviđa dovoljno dobro buduće događaje bez obzira što je T-test zaključio da nema značajnih razlika između vrijednosti.

5.2 Validacija i analiza modela predviđanja rezultata na mjesečnoj bazi

U ovom poglavlju korištena je nova tablica koja će nam prikazati mjesečna kretanja uvoza automobila s obzirom na vrstu goriva od 01.01.2010 do 31.05.2010.. Konfiguracija parametara prikazana je slikom 5-4 dok su rezultati predviđanja prikazani na slici 5-5.

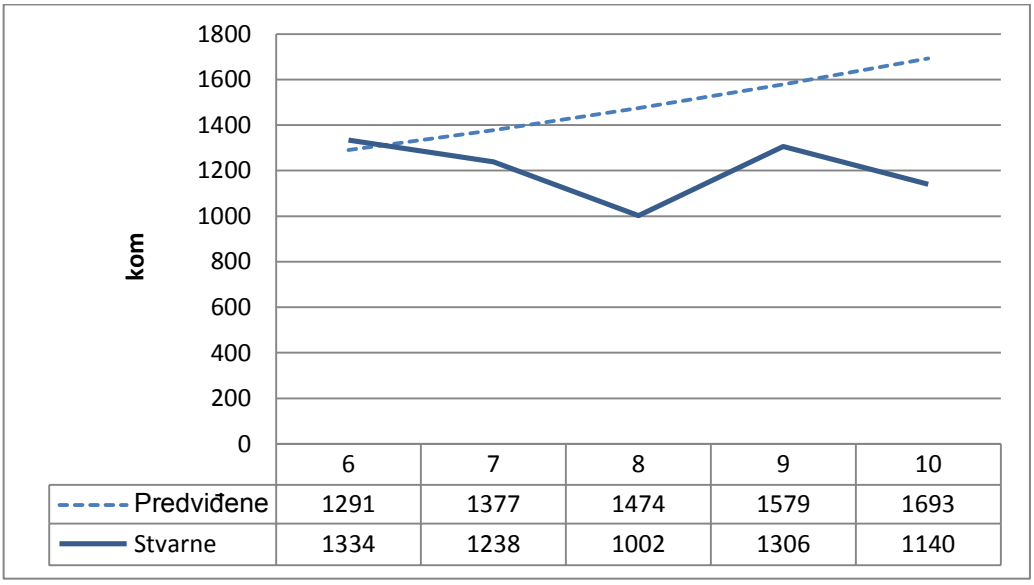


Slika 5-4. Konfiguracija parametara



Slika 5-5. Rezultati predviđanja

Slika 5-6 pobliže prikazuje stvarne i predviđene vrijednosti.



Slika 5-6. Graf stvarnih i predviđenih vrijednosti

Rezultati analize prikazani su tablicom 5-3.

Tablica 5-2. Rezultati analize

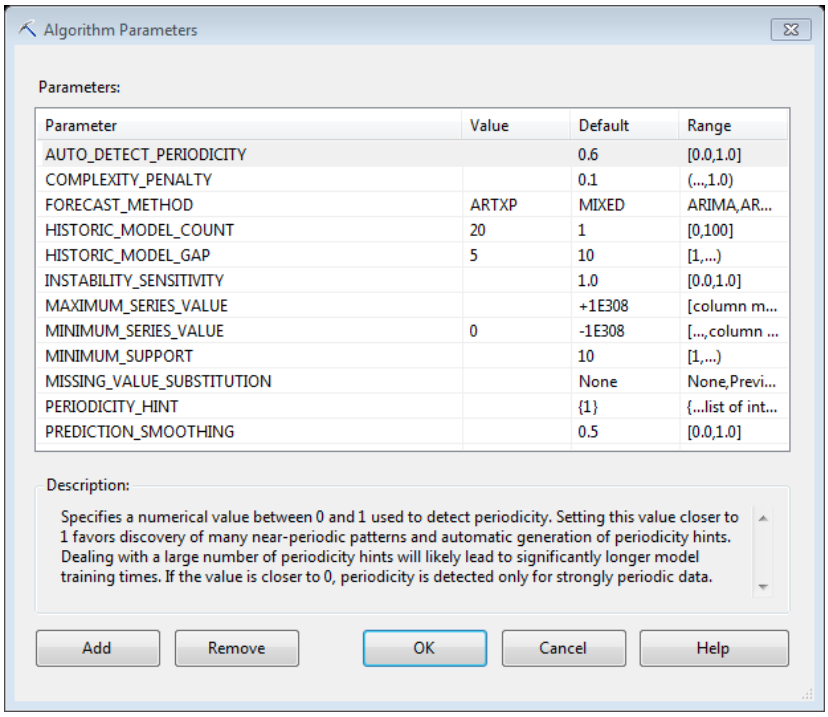
	Stupanj korelacije r	*P vrijednost t-testa (α=0,05)*
Stvarne vrijednosti / **Predviđene vrijednosti**	- 0,352	0,0175

Na osnovu podataka iz tablice 5-3 vidljivo je da je vrijednost P manja od 0,05 te se na temelju T-testa može odbaciti nultu hipotezu i zaključiti da se predviđene i stvarne vrijednosti značajno razlikuju. Koeficijent korelacije iznosi $r = -0{,}352$. Ovaj model nije dobar za predviđanje budućih podataka. Zašto? Premalo podataka za treniranje!

Pošto niti jedan set prethodnih podataka nije zadovoljavao predviđanja, ili zbog konfuznog broja podataka ili pak premalog, sljedeće setove podataka ćemo podijeliti i trenirati na tromjesečjima. Ova baza podataka, kako je prije navedeno, sastoji se od podataka za ukupno 10 mjeseci, stoga će se analizirati tri modela i to redom za prvo, drugo i treće tromjesečje. U sljedećim poglavljima bit će ubačeni dodatni podaci u bazu podataka i to na način da se ubace datumi sa nedjeljama i pripadne vrijednosti količina koje nedjeljom iznose 0.

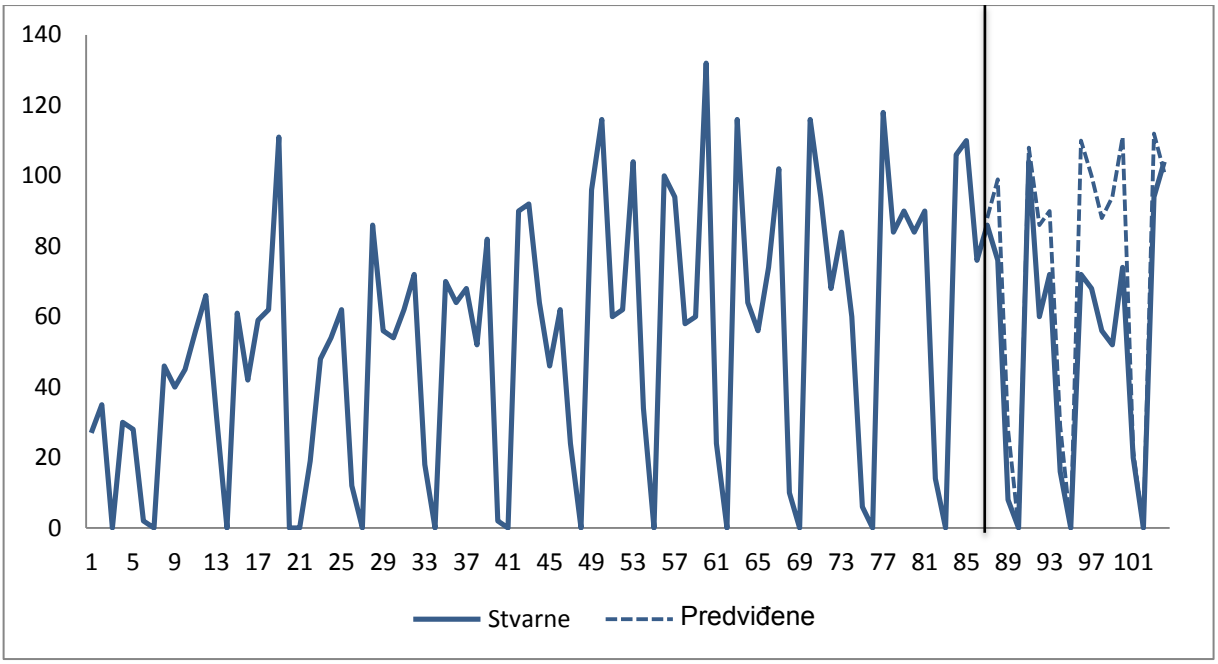
5.3 Validacije modela za prvo tromjesečje

Za ovaj model korišten je set podataka koji je treniran tri mjeseca (prvo tromjesečje) te su na temelju treninga dana predviđanja za sljedećih 18 dana. Na slici 5-7 je prikazana konfiguracija parametara za dobivene rezultate (slika 5-8).



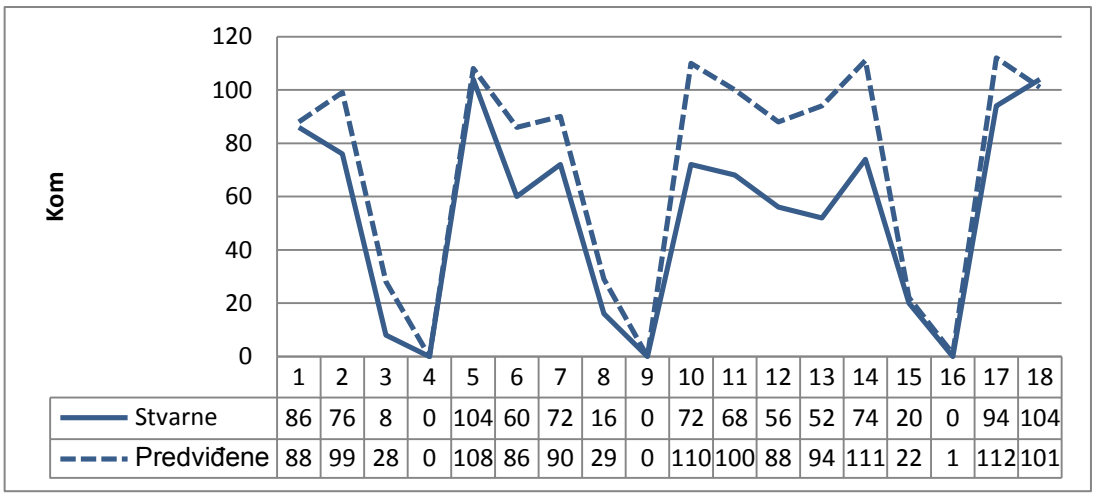
Slika 5-7. Konfiguracija parametara

Za ovaj model korišten je ARTXP algoritam koji prema zadanim vrijednostima daje bolje rezultate za kratkoročna predviđanja. Konfiguracija parametara je bitan segment u modelima predviđanja te se koristi kao iterativan postupak kako bi se došlo do željenih rješenja.



Slika 5-8. Rezultati predviđanja

Na slici 5-9 prikazani su rezultati predviđanja za sljedećih 18 dana zajedno sa stvarnim vrijednostima te je na njima naknadno izvršena statistička analiza.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Stvarne	86	76	8	0	104	60	72	16	0	72	68	56	52	74	20	0	94	104
Predviđene	88	99	28	0	108	86	90	29	0	110	100	88	94	111	22	1	112	101

Slika 5-9. Graf stvarnih i predviđenih vrijednosti

Nad ovim podacima provedena je statistička analiza u vidu T testa i koeficijenta korelacije. U tablici 5-4 prikazani su rezultati provedene analize koji preko P vrijednosti upućuju na to da nema značajne razlike između podataka.

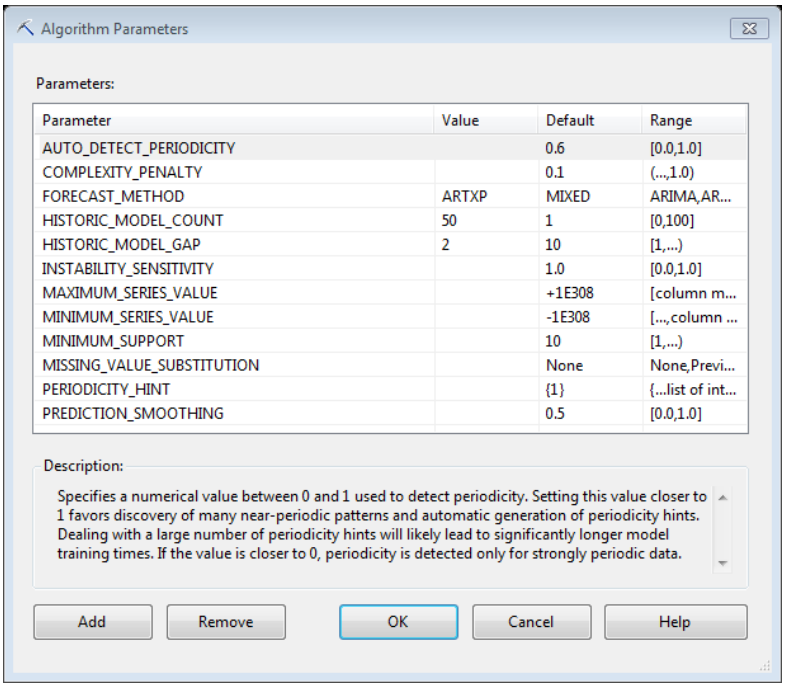
Tablica 5-3. Rezultati analize

	Stupanj korelacije r	*P vrijednost t-testa (α=0,05)*
Stvarne vrijednosti	0,939	0,212
Predviđene vrijednosti		

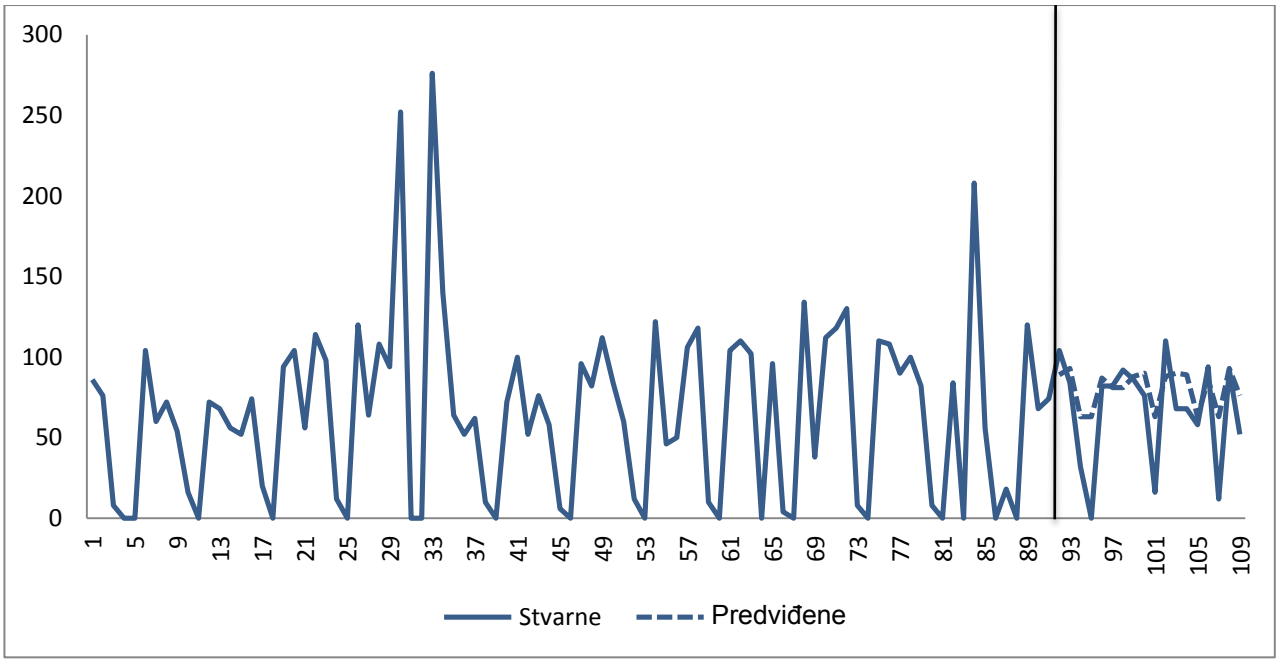
Prema dobivenom koeficijentu korelacije može se zaključiti da predviđene vrijednosti jako dobro slijede stvarne vrijednosti. Uzimajući u obzir i T test te koeficijent korelacije dobili smo cjelovit predviđajući model gdje imamo i zadovoljavajuću P vrijednost gdje je P = ~0,212 (P = 0,212>0,05; prihvaća se nulta hipoteza – nema značajne razlike između podataka) te koeficijent korelacije r = ~0,939 ($r \geq 0,7$ je jako dobra korelacija). Očigledno ovaj model zadovoljava oba uvjeta te se proglašava reprezentativnim za predviđanje budućih događaja. U konačnici, potreban je optimalan broj podataka za učenje, podešavanje parametara prema periodičnosti tih podataka te željeni vremenski faktor o tome koliko daleko želimo predviđati. Neki modeli, kao ovaj, će zadovoljavati predviđanja i na malom broju treniranih podataka.

5.4 Validacija modela za drugo tromjesečje

Za ovaj model korišten je set podataka koji je treniran tri mjeseca (drugo tromjesečje) te su na temelju treninga prikazana predviđanja za sljedećih 18 dana. Slika 5-10 prikazuje konfiguraciju parametara za dobivene rezultate (slika 5-11).

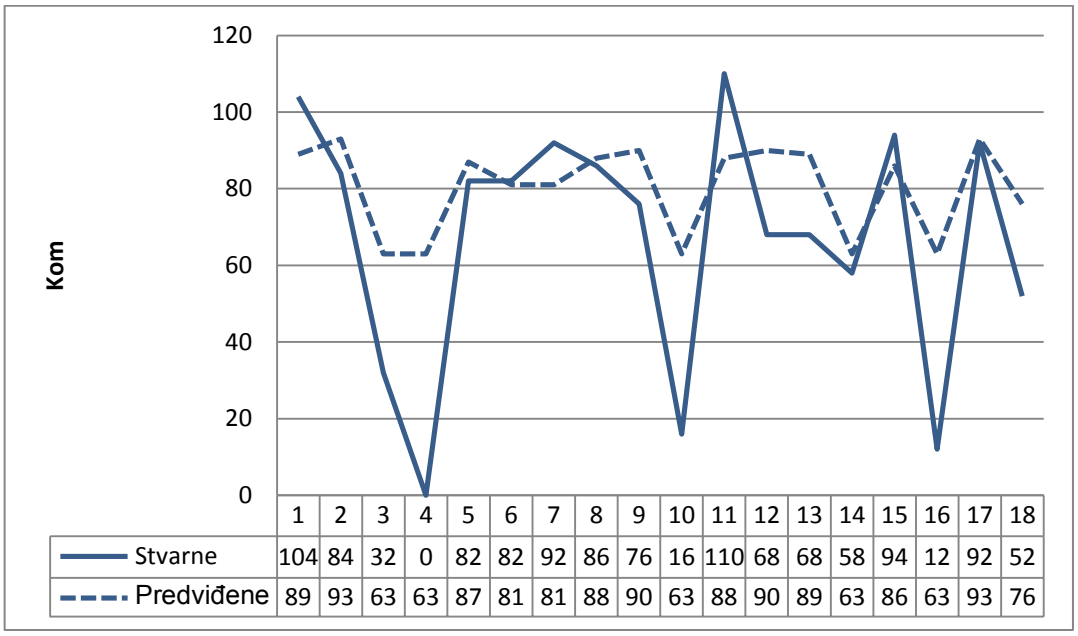


Slika 5-10. Konfiguracija parametara



Slika 5-11. Rezultati predviđanja

Slika 5-12 pobliže prikazuje dobivene predviđene u odonsu na stvarne vrijednosti.



Slika 5-12. Graf stvarnih i predviđenih vrijednosti

Sljedeće što preostaje jest provesti statističku analizu izračunom P vrijednosti i stupnja korelacije. Rezultati statističke analize su prikazani u tablici 5-5.

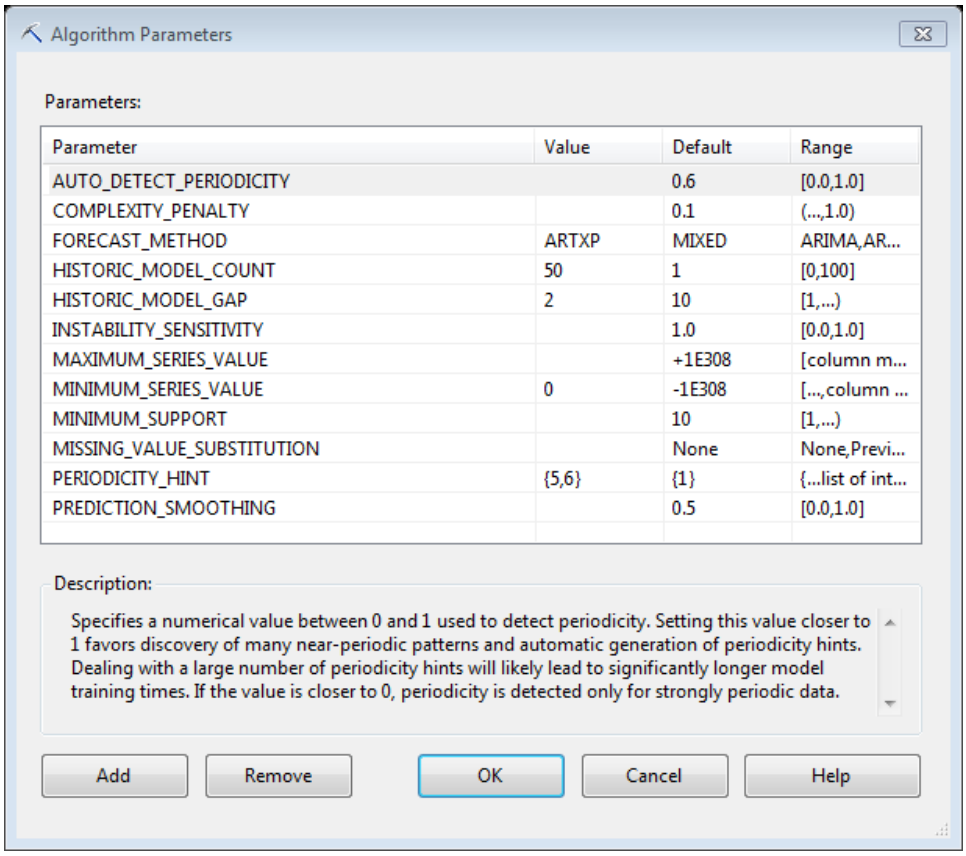
Tablica 5-4. Rezultati analize

	Stupanj korelacije r	***P vrijednost t-testa (α=0,05)***
Stvarne vrijednosti **Predviđene vrijednosti**	0,846	0,120

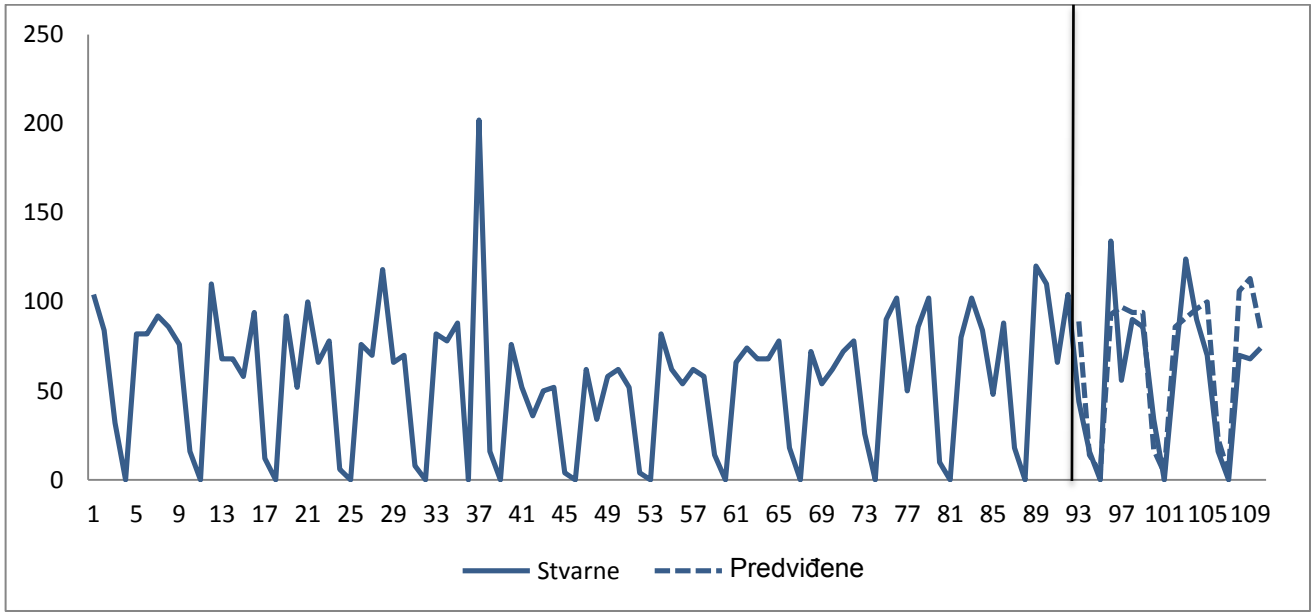
Zaključujući prema statističkoj analizi i ovaj model približno zadovoljava predviđanje budućih vrijednosti. Stupanj korelacije *r* je veći od 0,7 i P vrijednost je veća od 0,05.

5.5 Validacija modela za treće tromjesečje

Za ovaj model korišten je set podataka koji je treniran tri mjeseca (treće tromjesečje) te su na temelju treninga prikazana predviđanja za sljedećih 18 dana. Slika 5-13 prikazuje konfiguraciju parametara za rezultate predviđanja prikazane slikom 5-14.

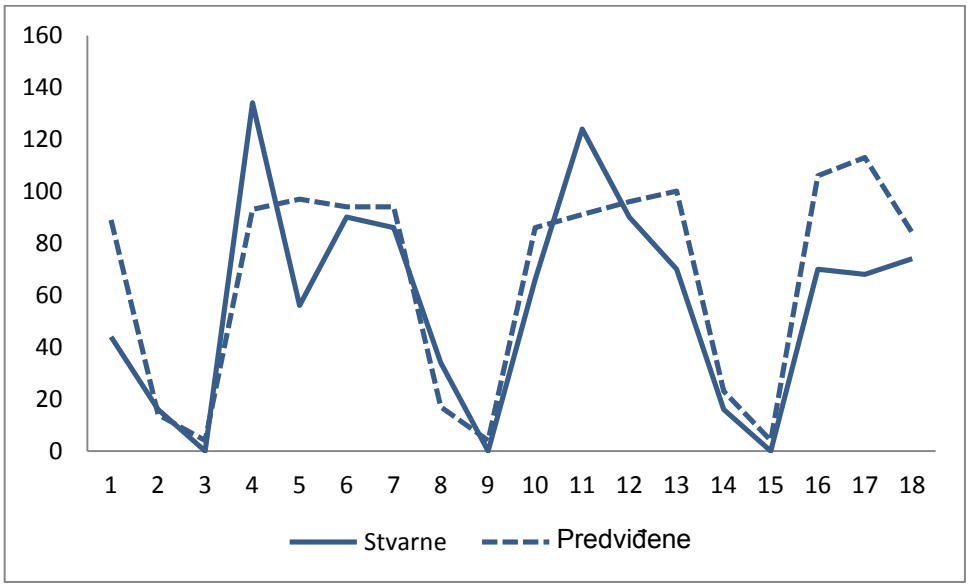


Slika 5-13. Konfiguracija parametara



Slika 5-14. Rezultati predviđanja

Pobliže prikazani rezultati nalaze se na slici 5-15.



Slika 5-15. Graf stvarnih i predviđenih vrijednosti

Tablica 5-7 prikazuje rezultate statističke analize za prikazani model.

Tablica 5-5. Rezultati analize

	Stupanj korelacije r	*P vrijednost t-testa (α=0,05)*
Stvarne vrijednosti	0,825	0,493
Predviđene vrijednosti		

Iz tablice 5-7 može se vidjeti da je stupanj korelacije zadovoljavajući i P vrijednost daleko veća od vrijednosti 0,05 te se zaključuje da i ovaj model dobro predviđa buduće vrijednosti.

6. Zaključak

U diplomskom radu su ukratko opisani aspekti poslovne inteligencije, kao i važnost donošenja učinkovitih i pravovremenih poslovnih odluka. Detaljno su opisane faze implementacije rudarenja podataka u sustav odlučivanja poduzeća. Faze uključuju razumijevanje vrste sadržaja i korištenje atributa kao izvora podataka. Pobliže je objašnjeno BIDS sučelje za rudarenje podacima kao i algoritmi odlučivanja.

Na stvarnom primjeru provedeno je rudarenja podataka u svrhu predviđanja trenda uvoza automobila s obzirom na vrstu goriva (Diesel). Na primjerima je Primjenjen je model *Microsoft Time Series* temeljen na dva programska algoritma ARTxp i ARIMA.

Analizirano je pet slučajeva predviđanja trenda uvoza vozila s obzirom na vrstu goriva u razdoblju od 10 mjeseci. U prva dva slučaja provedenom statističkom analizom utvrđeno je da se statistički podaci ne poklapaju tj. modeli nisu dovoljno dobro opisivali buduće događaje. U tom smjeru podaci su raspodjeljeni na tri seta podataka, odnosno na tri tromjesečja koja su podvrgnuta analizi. Prilagođavanje baze podataka za provedbu slučaja predviđanja za prvo tromjesečje rezultiralo je zadovoljavajućim statističkim vrijednostima, P = ~0,212 (P = 0,212>0,05; prihvaća se nulta hipoteza – nema značajne razlike između podataka) te koeficijent korelacije r = ~0,939 ($r \geq 0,7$ je jako dobra korelacija). Isto tako, uočeni je da pravilnim i iterativnim podešavanjem parametara utječemo na preciznost predviđajućih modela, no to zahtjeva i višu razinu poznavanja software-a. Sljedeća dva slučaja su također zadovoljavala statističke uvjete pošto su im se poklapali dobiveni statistički podaci. Statistički rezultati svih slučajeva prikazani su tablicom 6-1.

Tablica 6-1. Usporedba statističkih rezultata

	Stupanj korelacije r	P vrijednost t-testa (α=0,05)
Sedam mjeseci	- 0,304	0,172
Mjesecna baza	- 0,352	0,0175
Prvo tromjesečje	0,939	0,212
Drugo tromjesečje	0,846	0,120
Treće tromjesečje	0,825	0,493

U konačnici, korištena baza podataka nije reprezentativan primjer primjene modela za predviđanje zbog vremenske ograničenosti na samo 10 mjeseci. U tom smjeru, reprezentativnija bi bila baza podataka sa višegodišnjim sadržajem gdje bi se na primjeru tromjesečja uspoređivali podaci istih tromjesečja sljedećih godina. Takvim pristupom, iz godine u godinu, obogaćivali bi se kvartalni setovi podataka koji bi s vremenom postajali sve homogeniji u svrhu poboljšanja predviđanja. Na današnjem suvremenom tržištu potrebno je i od velike je važnosti sakupljati podatke i informacije kao i petvoriti ih u znanje stalnim analizama i traženjem skrivenih veza između pojedinih, naočigled, nepovezanih varijabli, a sve u svrhu racionalizacije poslovanja, predviđanja i preciznijeg odgovora na zahtjeve klijenata.

7. Literatura

[1] Lynn: Smart Business Intelligence Solutions with Microsoft SQL Server 2008, Microsoft Press, 2009

[2] Vercellis, Carlo: Business Intelligence: Data Mining and Optimization for Decision Making, John Wiley & Sons, 2009 Langit

[3] WEB: MSDN Library; http://msdn.microsoft.com/en-US/, 2012.

[4] WEB: Microsoft SQL Server Library; http://technet.microsoft.com/en-us/library/bb545450(v=msdn.10%20).aspx, 2012.